Reusing Ontologies to Enrich Semantically User Content in Web2.0: A Case Study on Folksonomies

Thesis

UNRESTRICTED.

Reusing Ontologies to Enrich Semantically User Content in Web2.0: A Case Study on Folksonomies

Sofia Angeletou
Dipl. Engineer of Computer Science and Informatics

A Thesis presented for the degree of
Doctor of Philosophy

Semantic Web and Knowledge Services
Knowledge Media Institute
The Open University
England

September 2010

Date of Submission: 30 September 2010
Date of Award: 22 February 2011

To my parents
Petroula and Babis
and to my sister
Arxontoula

Reusing Ontologies to Enrich Semantically User Content in Web2.0: A Case Study on Folksonomies

Sofia Angeletou

Submitted for the degree of Doctor of Philosophy

September 2010

Abstract

Semantic Web and Web2.0 emerged during the past decade promising to achieve new frontiers for the Web. On the one hand, the Semantic Web is an interlinked web of data, supported by ontological semantics and allowing for intelligent applications such as semantic search and integration of heterogeneous content across systems and applications. On the other hand, Web2.0 represents the new technologies and paradigms that revolutionised the user engagement in content creation and introduced novel means towards social interaction. Bridging the gap between Web2.0 and the Semantic Web has been proposed as a means to better manage and interact with the large amounts of user contributed content, which is a new challenge for Web2.0. This thesis focuses on a popular paradigm of Web2.0, folksonomies. In particular, we investigate the semantic enrichment of folksonomy tagspaces by reusing ontologies available in the Semantic Web. We identify the need for methods that automatically apply semantic descriptions to user generated content without requiring user intervention or alteration of the current tagging paradigm. We use an iterative approach in order to identify the characteristics of folksonomies and the attributes of knowledge sources that influence the semantic enrichment of tagspaces. We build on the results of our experimental studies to implement a folksonomy enrichment algorithm, that given an input tagspace, automatically creates a semantic structure that describes the meaning and relations of tags. We introduce measures for the evaluation of enriched tagspaces and finally, we propose a search algorithm that exploits the semantic structures to improve folksonomy search.

Acknowledgements

"*Neo, sooner or later you're going to realise just as I did that there's a difference between knowing the path and walking the path*" Morpheus, The Matrix, 1999.

There are many people who walked with me on the path to this PhD without whose help, love and support I would have never made it.

First and foremost I want to express my gratitude to my supervisors Enrico and Marta for believing in me and giving me the opportunity to make all this happen.

Being next to Enrico for four years taught me more than research skills, it taught me the value of common sense, the value of research ethics and the value of a winner's attitude. Enrico has this unique ability to make his point using all kinds of metaphors such as "*you can't build this supercool car assuming it will work fine with three wheels, and then go back and say, ops! I need another one*". Thank you for being my supervisor, for teaching me valuable lessons and for sacrificing your tennis sessions to read my chapters!

Marta. There are really no words to describe her and what she means to me. I was supervised by one of "AI's ten-to-watch" and although I could easily be intimidated, Marta's behaviour didn't allow me to. From the very beginning she treated me as a colleague and as a friend. She taught me how to do research, how to manage my research, how to write papers, how to make presentations and have a life at the same time. I cannot think of a single task which I complete without recalling her words of wisdom and her advice. I haven't met anybody with the professionalism, drive and success of Marta who is, at the same time, as humble and warm hearted as her. Marta, I cannot thank you enough for teaching me, helping me and sharing my path for the last four years. I knew we would part at some point in our careers but I miss you very very very much.

Last, but not least of my supervision team I want to thank my external supervisor Lucia. We didn't manage to spend as much time together as I would have liked to but your help and feedback have been very useful to my work. Thank you.

I want to thank all the people who contributed to my research with their time and valuable comments. The partners of the NeOn project for spending their lunch breaks with me discussing my work and giving me helpful suggestions, the evaluators who participated in my user studies and helped me get my results and all the colleagues I met in conferences and other events and exchanged thoughts and ideas. I want to thank Mathieu for helping me clarify my algorithm and set up my evaluation strategy and Vanessa for sharing with me her server and her salsa nights! (If you think this is irrelevant to research, you may want to reconsider). Finally, I would have never done it without the help of Laurian who was always there with a great and quick solution to my tons of technical problems. His help has been catalytic in the achievement of many deadlines. Laurian, thank you and although I never understood your love for weird food I miss having you around (there's nobody to talk to me about century eggs!)

During this time KMi has been more than a workplace for me it has been my home. Apart from being a great place to work, the support of KMi-ers has been exceptional. I want to thank the administration team, Aneta, Ortenz and Jane for making my life easier dealing with all kinds of small and big issues, the IT team Damian, Lewis, Robbie and Paul for giving me technical support faster than I could have ever dreamt of and the graphic designers Harriet and Peter for visualising my ideas.

I want to acknowledge the support and encouragement of Harith during the last months of my PhD. Thank you for believing in me and for giving me something exciting to look forward to.

My wonderful friends coloured my life in Milton Keynes, supported me in times of gloom and helped me carry on. Carleto, La Bella Gemma, Rosaria, Carlitos, Daviduccio and Gary with whom I have shared some great moments to remember and I am looking

forward to more!

There are two ladies who I wanted to thank along with my friends, but I don't really consider them friends, I consider them family. Actually, they are two "pesadas" who have made my life impossible with their constant fights over the best olive oil! Yes, I consider them family, my two guardian angels, who have walked this path with me and have always been there to celebrate my happiness and wipe my tears. Annalisa and Ainhoa, tesorino and amorcito, I cannot even think how I would have survived here if I hadn't met you. Thank you for joining my path and making Milton Keynes feel like home (!!!). My Barese angel, your shouting is incomparably effective. I owe you, Big Time!

I want to close this section talking about the most important people in my life, which I have missed the most during the past four years, my family. The people who made me who I am. The people who taught me what is right and what is wrong and gave me the principles to be a good person. My parents Petroula and Babis who have given me all the love in the world and have worked really hard to give me the opportunity to create my own paths. My sister, Arxodoula, who is always on my side, on any path that I choose, holding my hand and reminding me what is this all about. If I could select the best family I would only select you. I love you.

Publications

- Sofia Angeletou, Marta Sabou and Enrico Motta, (2009) *Improving Folksonomies using Formal Knowledge: A Case Study on Search*, Proceedings of The 4th Asian Semantic Web Conference, Shanghai, China.

- Sofia Angeletou, Marta Sabou and Enrico Motta, (2009) *Improving Search in Folksonomies: A Task Based Comparison of WordNet and Ontologies*, Poster at The 5th International Conference on Knowledge Capture, Redondo Beach, California.

- Sofia Angeletou, Marta Sabou and Enrico Motta, (2009) *Folksonomy Enrichment and Search*, Demo at The 6th European Semantic Web Conference, 6th European Semantic Web Conference, Crete, Greece.

- Sofia Angeletou, (2008) *Semantic Enrichment of Folksonomy Tagspaces, International Semantic Web Conference*, Doctoral Consortium of The 7th International Semantic Web Conference, Karlsruhe, Germany.

- Sofia Angeletou, Marta Sabou and Enrico Motta, (2008) *Semantically Enriching Folksonomies with FLOR*, 1st International Workshop on Collective Semantics: Collective Intelligence & the Semantic Web (CISWeb 2008) at The 5th European Semantic Web Conference, Tenerife, Spain.

- Mathieu d'Aquin, Enrico Motta, Marta Sabou, Sofia Angeletou, Laurian Gridinoc, Vanessa Lopez and Davide Guidi, (2008) *Towards a New Generation of Semantic Web Applications*, IEEE Intelligent Systems, 23, 3, pp. 20-28.

- Marta Sabou, Jorge Gracia, Sofia Angeletou, Mathieu d'Aquin, and Enrico Motta, (2007) *Evaluating the Semantic Web: A Task-based Approach*, The 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference, Busan, Korea

- Sofia Angeletou, Marta Sabou, Lucia Specia and Enrico Motta, (2007) *Bridging the Gap Between Folksonomies and the Semantic Web: An Experience Report*, Workshop on Bridging the Gap between Semantic Web and Web 2.0 at The 4th European Semantic Web Conference, Innsbruck, Austria.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Background and Motivation

Folksonomies [122], are typical Web2.0 [92] applications used to publish, annotate and share content on the web. They are highly popular due to their "*low entry barriers*" and effectiveness in personal and network organisation [84]. They are approachable by the casual web user by not requiring the latter to have special skills or technical knowledge. For content annotation, also called *tagging*, there is no controlled vocabulary or other syntax rules and **no constraints are imposed on the users**. While liberal tagging is the main advantage of folksonomies, it also introduces certain limitations. For example, the different backgrounds and expertise of the users on various topics are reflected in their selection and usage of tags. As a result, users with different vocabularies are not likely to encounter each other's content, unless the content semantics are considered by the tools supporting user navigation. In particular, phenomena such as polysemy, synonymy and basic level variation [51] may hamper the performance of folksonomies in terms of content retrieval, content recommendation and so on, thus limiting the user experience. In addition, the rich user contributed content, cannot lend itself easily to applications that require some sort of intelligent integration of information. As a result

it only remains available to the closed specific applications where it was created.

In contrast to the low effort required to annotate folksonomy resources, the creation of ontologies is a laborious process. Ontologies [53] are knowledge artefacts used to formally specify objects, their "behaviours" and their relations in the context of various domains, tasks and applications. For their creation, knowledge engineers and domain experts are required to collaborate in order to produce **technically and conceptually sound models**. Ontologies are considered the backbone of the Semantic Web and play a key role in its realisation [29].

Both academic researchers [28, 33, 52, 58, 70, 110] and corporate stakeholders [23, 67] have identified the need for semantics in order to solve the problems of Web2.0 and achieve new frontiers on the web. Baeza-Yates et. al [24] discuss the merits of semantic metadata towards new paradigms of search and Benjamins et. al [28] highlight the need for semantics as a means to manage and organise the "*vast content of Web2.0*". Alani [20] states that "*a semantically-enabled content-exchange channel offers direct benefits with respect to consistency checking, relative ease of integration and distributed querying, and efficient data and information exchange and merging*".

Hendler and Golbeck [58] describe the alignment of tags to semantics as a first step towards achieving prospects such as interlinking items, users, networks and communities on the web and addressing content retrieval and integration issues. These prospects cannot be achieved only by over-annotating the content if this still lacks a semantic description [57]. Hence, **an immediate challenge is the application of semantics to tags**. At the same time, the maturing of Semantic Web technologies has allowed for the creation and publication of ontologies on the web. Hendler and Golbeck [58] claim that the creation of correct ontologies is laborious but they can have a major impact on tagging systems. d'Aquin et. al [40] endorse the reuse of existing knowledge encoded in ontologies to solve classic problems, such as ontology matching and question answering, as well as addressing the issues of Web2.0 and folksonomies. As a result an

additional challenge emerges **considering the reuse of existing knowledge.**

1.2 Research Problem

The research carried out in the scope of this thesis addresses the two aforementioned challenges. In particular we investigate **how and to what extent the user tags in folksonomies can be semantically enriched by reusing existing semantics** and **what are the benefits of such enrichment**? This problem can be further specified:

RQ1: To what extent can folksonomies' tagspaces be semantically enriched by automatically exploiting semantic structures built in online ontologies? Answering this question requires one more level of analysis and is represented as follows: How can we discover automatically the meaning of individual tags and the semantic relations between tags based on their context? How can existing ontologies be exploited for the enrichment of tags?

RQ2: What other resources are required in case the Semantic Web falls short of this task? In case the usage of ontologies is not sufficient to semantically enrich tagspaces, what other sources should be exploited for the purposes of folksonomy enrichment.

RQ3: How can the enriched tagspaces and enrichment processes be evaluated? What measures and evaluation strategies should be established to quantify the performance of the enrichment methods?

RQ4: How can the enriched tagspaces be exploited and evaluated in the context of content retrieval? What methods should be created for improving folksonomy search utilising the enriched tagspaces? What measures should be established to asses the value of enriched tagspaces in search?

In the following section we describe the course of our research motivated by these questions.

1.3 Methodology and Structure of the Thesis

Due to the open, heterogeneous and dynamic nature of the resources involved in our approach, both ontologies and folksonomies, we performed a series of exploratory studies prior to establishing the requirements for the final enrichment and search algorithms. We used real world data to test various hypotheses and the results of our investigations were exploited for the implementation of our final approach. Next we briefly introduce the work described in the following chapters and highlight the outcomes obtained by each study and how they motivated the next steps of our work.

In **Chapter 2** we present a review of the existing work on folksonomies. We identify the most popular research lines and give an overview of the most important studies presented in each. Finally, we identify the open issues on the folksonomy research area and position our work.

In **Chapter 3** we present an analysis of the entities of folksonomies, their relations and the issues which influence search and content organisation. We define the core objects of folksonomies and introduce the concepts we use through out this thesis to describe the semantic enrichment of tagspaces. We present the schema of the ontology we built to support the enriched tagspaces. Finally, we introduce a set of measures for the evaluation of the enrichment algorithms, the semantic structures and their influence on search.

Our first attempt to automatically enrich tagspaces was presented in [21]. We reused the clusters generated by Specia and Motta [114] and applied the relation discovery algorithm implemented by Sabou et. al [108]. In that work we automatically enriched tag clusters and obtained useful insights on the types of tags and their relations, as well

as, on the characteristics of Knowledge Sources that influence the enrichment process.

With the first version of FLOR, FLOR-1, presented in **Chapter 4** we aimed to increase the coverage of tags from ontologies (compared to [21]) by utilising an up-to-date ontology search mechanism. In addition, we used WordNet to expand tags with their synonyms before seeking potential defining entities in ontologies. The major outcome of the experiments carried out with FLOR-1 was the adverse impact of WordNet (as a source for disambiguation and semantic expansion) in combination with hierarchical similarity measures on the enrichment process (L4.3). The identification of non hierarchical relations among the tags lead to the conclusion that the alternative relatedness measures are required (L4.1). In addition, we discovered the need for statistical measures where semantic measures fail (L4.2).

With the study presented in **Chapter 5** we aimed to obtain user incentives on semantically-enabled search and identify additional issues of the FLOR-1 enrichment algorithm that are only projected during search. We used a domain restricted dataset from Flickr, in order to guide the user queries. Unfortunately, using the same version of the enrichment algorithm (FLOR-1) that exploits WordNet to disambiguate and expand the tags once again led to poor coverage of tags by ontological entities. Therefore, we conducted the experiment relying only on the WordNet derived hierarchy of tags. Despite its negative effect on the tag anchoring to semantic entities, WordNet's usage as a Knowledge Source for the creation of sense structures for search was successful (L5.3). This prompted us to reconsider its usage and instead of employing it as a disambiguation and semantic expansion resource to use it as a Knowledge Source for enrichment. In terms of user experience, the participants of our experiment reported that the organisation of results using semantically-enabled search was meaningful and helped them generate ideas for query reformulation (L5.4). We also identified the importance of the tagspace coverage from a semantic entity's neighbours in the enrichment value of the semantic entity (L5.1). Finally, we validated the need for statistical relatedness measures (L4.1, L4.2) in order to exploit non-semantically related contexts

(L5.2).

In an effort to compare the impact of WordNet and ontologies in terms of creating sense structures and exploiting them for search we performed a Knowledge Source comparison study in **Chapter 6**. Indeed, we noted that the value of WordNet is comparable to the value of ontologies and decided to use it as a Knowledge Source for enrichment rather than expansion (L6.1). The difference is that in the second case, when a tag is correctly assigned to a WordNet synset we consider the tag enriched and terminate the process, provided that there are no other ontologies containing appropriate definitions. In Chapter 6 we also compared semantically-enabled search with cluster-based search from folksonomies. We observed that although cluster-based search catered for idiosyncrasies and returned less groups (L6.2), the semantically-enabled search presented the results in meaningful categories (L6.3). Finally, we discovered that failure of integration that leads to different senses with the same meaning, has an adverse impact on search by generating overlapping groups (L6.4).

Chapter 7 presents one of the core contributions of this work, which is the folksonomy enrichment algorithm FLOR-2. We describe how the results of the previous studies are transformed into requirements, which are then used as a basis for the design and development of the algorithm. We detail the individual phases and steps of the algorithm and how each contributes to the final output. The main improvements of FLOR-2 compared to FLOR-1 were focused on the exploitation of WordNet as a Knowledge Source, and on the phases of sense disambiguation, sense integration and semantic aggregation.

In **Chapter 8** we evaluate FLOR-2 and validate its improved performance compared to FLOR-1 in terms of sense assignment correctness and tagspace coverage. We do so by enriching the same dataset with the two versions of the algorithm and contrasting the results. We evaluate further the performance of the improved enrichment algorithm using an additional dataset and assessing the degree of connectivity of the semantic structure created by FLOR-2. Finally we present a quantitative and qualitative analysis

of the enrichment process and the impact of the design decisions on the output of the algorithm.

In **Chapter 9** we present a query algorithm, exFLORe, which exploits the enriched tagspaces in order to improve search in folksonomies. exFLORe is based on the outcomes of the experimentations presented in chapters 5 and 6 concerning the influence of semantic structures on search (L5.4, L6.2, L6.3). This search algorithm maps query keywords to senses in the enriched tagspace, disambiguates the senses and returns the results associated with them in a ranked order.

Chapter 10 concludes the thesis with an overview of the work carried out, the contributions of this study and an outlook for future work.

1.4 Contributions

An algorithm for the enrichment of folksonomy tagspaces. We present an algorithm that semantically enriches folksonomy tagspaces by explicitly assigning meaning to tags and describing their inter-relations. Our method for folksonomy enrichment adheres to the following principles:

1. **Domain Independent.** The process does not assume domain restriction, neither in the selection of ontologies nor in the selection of tags.
2. **Automatic.**
   - The enrichment process operates on existing tagspaces, does not require user feedback during the tagging activity and does not suggest a shift in the tagging paradigm.
   - There is no need for preselection of Knowledge Sources, all ontologies available online are considered.
3. **Uses Heterogeneous Knowledge.** Our algorithm integrates knowledge from online available ontologies and WordNet.

4. **Unsupervised.** There is no need for training data

5. **Creates an explicit semantic structure.** The output of the algorithm is an explicit semantic structure supported by an appropriate ontology.

Evaluation measures for the semantic structures that represent the tagspaces and for the enrichment algorithm. With these measures we evaluate the connectivity and richness of the semantic structures and the performance of the algorithm in terms of tagspace coverage.

Search algorithm for enriched tagspaces that exploits the enriched tagspaces to improve search by addressing the issues of polysemy, synonymy and basic level variation and allows for result diversification.

Chapter 2

Literature Review on Folksonomies

In this chapter we present a review of the relevant work that aims to understand and improve folksonomies. We analyse the most significant and representative approaches proposed to date and present a goal-based overview of the literature according to the most widely investigated folksonomy problems. Finally, we summarise the existing work and highlight the open issues addressed by this thesis.

2.1 Introduction

The success of Web2.0 has motivated a broad line of investigations on understanding and improving this type of user generated content. Considering that the contribution of users was estimated to be four to five times larger than that from the professional publishing on the web[1] [104], the need for organisation and management tools emerged rapidly. The vision that using semantics to describe, organise and exploit such vast amounts of data would bring a new era on the web [28, 33, 52, 58, 70, 110] motivated a broad range of initiatives.

[1]Ramakrishnan et.al estimate that in 2007 users generated 8 to 10GB of content while at the same time only 2GB was generated by the professional web.

A successful example of user contribution on Web2.0, Wikipedia [14], was used in one of the first semantification initiatives with the aim to make the knowledge in it available on the web of data. The effort [31] to create a semantic version of Wikipedia resulted in DBpedia [6], a well adopted knowledge resource. Additional works towards the same vision describe how the usage of semantic web technologies can improve knowledge sharing, enable novel paradigms in online communities [32, 38, 39, 50, 71, 113], and outline the architectural integration of Semantic Web and Web2.0 [91, 93].

This thesis focuses on the special case of folksonomies, also called tagging systems, and for this reason we limit the literature analysis to work which focuses on them. This research area is quite young and there are no established methodologies, approaches and practices. In addition, classifying existing work in the area is a complex task due the multidimensionality of tagging systems and the interdisciplinary diversity of proposed approaches. Nevertheless, in the following, we aim to give a comprehensive overview of the literature from two different perspectives.

Due to the young age of the domain, a great amount of work has been published aiming at **understanding tagging systems**. In particular, investigations on the incentives and behaviour of users, the types and characteristics of tags, and the network dynamics of folksonomies are very important to understand and establish the field. We analyse the outcomes of these studies in Section 2.2.

The open, dynamic nature and the inherent problems of folksonomies presume many **different possibilities towards improving tagging systems** (see Section 2.3). For the sake of providing a detailed overview, as well as being consistent with the goals of this work, we present the folksonomy improvement work according to two abstract (and not totally mutually exclusive) groups. In the first group we discuss work that applies some sort of structure on folksonomies by utilising either emergent or explicit semantics. The second group includes work dealing with the improvement of content retrieval by enhancing search and content recommendation.

Finally, in Section 2.4 we summarise the existing work and discuss the open issues addressed by our approach.

2.2 Understanding Tagging Systems

The first observations and debates on tagging systems were published in weblogs [55, 56, 84, 99, 112, 123] as soon as folksonomies became widespread. There, the benefits as well as the drawbacks of free tagging were highlighted and compared to traditional annotation and classification schemes. Their main observations were that the cognitive effort in content annotation in folksonomies is much lower compared to classification, yet the lack of structure was highlighted as an impedance of folksonomies while classification schemes do not suffer of such lack of structure.

The first academic study on folksonomies was conducted by Golder and Huberman [51] in 2005, and its results influenced most of the research on the field[2]. They macroscopically studied the patterns and user incentives in folksonomies. They showed that the tagging activity follows a stable pattern after a certain amount of time and further validated their results via the visualisations provided by Cloudalicio.us [107]. Their major contribution, though, is a preliminary **categorisation of tags** based on the tagging motivation. More specifically, they identify tags used to denote the topic (`webdesign`), the type (`blog`), the owner (`timbl`) and various qualities of a resource (`cool`, `funny`). They also identify tags used for self reference (tags beginning with `my`) and task organising (`toread`). Their observations apply on data from Delicious [7], however, Zollers [136] obtained the same findings on different folksonomies (Amazon [1] and Last.fm [11]).

Marlow et. al [83] list **user incentives for tagging** (personal retrieval, contribution and sharing, attention seeking, self presentation, play and competition, opinion expres-

[2]According to Google Scholar, "*The structure of collaborative tagging systems*" was cited by 546 and the "*Usage patterns of collaborative tagging systems*" was cited by 637. (Accessed in April 2010)

sion) that affect the overall evolution and dynamics of folksonomies. They claim that tagging is also influenced by **system design decisions** (tagging rights, tagging support, aggregation model, object type, source of material, resource connectivity, social connectivity) while Diaz et.al [43] show how design decisions (keyword and spelling suggestions, display of the items annotations, ability to enter multiple keywords) also affect the user behaviour during search.

In terms of how **tagging behaviour** influences the global dynamics of the systems, Körner et.al [68] show that prolific taggers, who use a richer vocabulary and exhibit higher tagging verbosity, contribute better to the emergent semantics of folksonomies compared to the users who provide fewer tags. Fu et. al [48] show that the tagging behaviour of expert taggers converges faster than the behaviour of the novices possibly due to the ability of the expert taggers to better evaluate the topics of the resources and as a result assign more accurate and high quality tags.

Millen et. al [87] explore how the incentives of the users affect their **searching and browsing behaviour**. Employing a corporate folksonomy as a use case, they identify three types of search. The community browsing (popularity and recentness based navigation on community created bookmarks), the personal search (search on ones own bookmark collection), and the explicit search (traditional keyword search to locate bookmarks created by the community). They show that the most frequently engaged activity by the largest number of users in the system is the community browsing, followed by the explicit search. The most prolific taggers tend to engage more on personal search due to the fact that they use the system for personal organisation rather than sharing. More lightweight taggers, who spend less time tagging, use the community browsing and the explicit search to locate bookmarks created by the community. Exploiting social links to discover content has also been studied by van Zwol [120], who presented how the social connections affect the discovery and popularity of newly added items in Flickr.

From a **quantitative perspective** Al-Khalifa et.al [19] show that more than 60% of the tags represent common knowledge and can be used for content classification and metadata generation. 30% of the tags are personal, i.e., directed either to oneself or ones network, and around 4% express user opinion on the tagged resource. In terms of resource to tag ratio, Plangprasopchok et.al [100] show that users provide 4 to 7 tags per bookmark on Delicious while Rattenbury et. al [105] show that each photo on Flickr has an average of 3.74 tags. Finally, Bischoff et. al [30] support the annotation value of folksonomies by showing that more than 50% of tags in Flickr and Delicious bring new information to the resources while the same happens for more than 98.5% of the tags in Last.fm.

Another study that investigates the **annotation value of folksonomies** is presented by Kipp [66]. Using journal articles tagged in Citeulike [2] and Connotea [4] she comparatively explores keywords generated by the authors and by professional librarians and user tags from the above systems. She discovers that user tags are very valuable because they provide novel terminology and more generic terms. Al-Khalifa et.al [18] show that Delicious tags have a higher semantic value than the automatically extracted keywords from the text of the resource because they are assigned by users with broad backgrounds and variable expertise which may not be reflected in the text of the resource. On the opposite side, Lux et.al [78] show that there is a large number of tags that is inappropriate for retrieving resources or users. Such are misspellings, unpopular (infrequent) tags, shortcuts on resources, or personal vocabularies.

A different perspective on the **value of tagging systems** is given when studied under the prism of **web search improvement**. Tags are usually good summaries of the corresponding webpages in collaborative systems and their count indicates the popularity of webpages. In addition, the social interactions among the users provide novel interlinks among resources. Exploiting such network dynamics, modifications of the classic PageRank [94] algorithm have been tailored for the folksonomy network. Such are SBRank [127], FolkRank [60], SocialSimRank and SocialPageRank [25]. These algo-

rithms exploit the notions *tag and user popularity and similarity* to rank the resources and demonstrate the benefits of introducing the social dimension on the web. Along the same lines, Yeung et.al [130] show that the socially derived associations among resources can be used to improve the existing hyperlink structure of the web. Their evidence supports the fact that such user induced links are indeed of very high quality. In an earlier work [81] they deal with the problem of multiple and mixed results in the cases of ambiguous search terms. They approach this problem by diversifying web search results with classifiers based on tag co-occurrence from Delicious. Finally Mislove et. al [88] show that search based on social networks can be more efficient due to better disambiguation, serendipity and ranking of results.

Yeung et.al [128] study the **tag ambiguity problems** by introducing the concept of mutual contextualisation of entities in folksonomies. This means that one entity e.g., a tag, resource or a user is better understood when contextualised with the other two. On this contextualisation line, they [129] experiment with unsupervised clustering on different types of networks (tag-based document, user-based document, tag-context) and show that user-based networks, which encapsulate the social interaction element, perform better in tag disambiguation. Ronzano et. al use Wikipedia articles to build a sense repository called Tagpedia [106], which contains all possible candidate meanings for the tags. They then propose a disambiguation algorithm that exploits the co-occurrence of the tag to be disambiguated with the other tags in its context [116]. In this case the context of the tag consists of the other tags tagging the same resource enriched with the popular tags used globally to annotate the resource. Finally they assign each tag to a Tagpedia sense, and implicitly to a Wikipedia article and furthermore to a DBpedia entry. DBpedia is also used by Garcia et.al [49] who directly associate a tag to the possible DBpedia entries. For each entry a vector of related terms is extracted and compared to the vector describing the context of the tag using cosine similarity. Although the authors define different types of contexts such as resource-based, user-based, social-based and combinations of them, they only employ

the resource-based context, which is the co-occurring tags on the scope of a resource. The first approach, which is independent of external knowledge sources, can work for all the tags. However, the latter two approaches use resources with a high update frequency in new terminology, which increases the possibility of finding candidate senses for a tag. In addition, they provide explicit meaning for each tag by connecting it to a richly described sense.

2.3 Improving Tagging Systems

Motivated by the early realisation of, not only the benefits of free and unrestricted tagging, but also its adverse impact on content organisation and retrieval [51, 83], the research community dedicated a lot of effort in creating solutions which can alleviate such effects. In this section we present an overview of the work in this area from the viewpoint of the goals of this thesis. We first analyse how structuring folksonomies is approached in the literature (Section 2.3.1) and then discuss the approaches that enhance content retrieval in folksonomies (Section 2.3.2) according to the structure presented in Figure 2.1.

2.3.1 Structuring Folksonomies

The application of structure on user generated content has been highlighted [28, 33, 52, 58, 70, 110] as a crucial component towards realising the vision of an interoperable and intelligent web. Depending on their background discipline researchers have approached the application of structure to folksonomies in different ways. In this section we distinguish two types of approaches towards structuring folksonomies, the *implicit* and the *explicit*. The implicit way is to exploit their emergent semantics and discover relations among the elements of folksonomies. Such approaches usually yield hierarchical structures but the relations are not made explicit. On the contrary, the explicit way

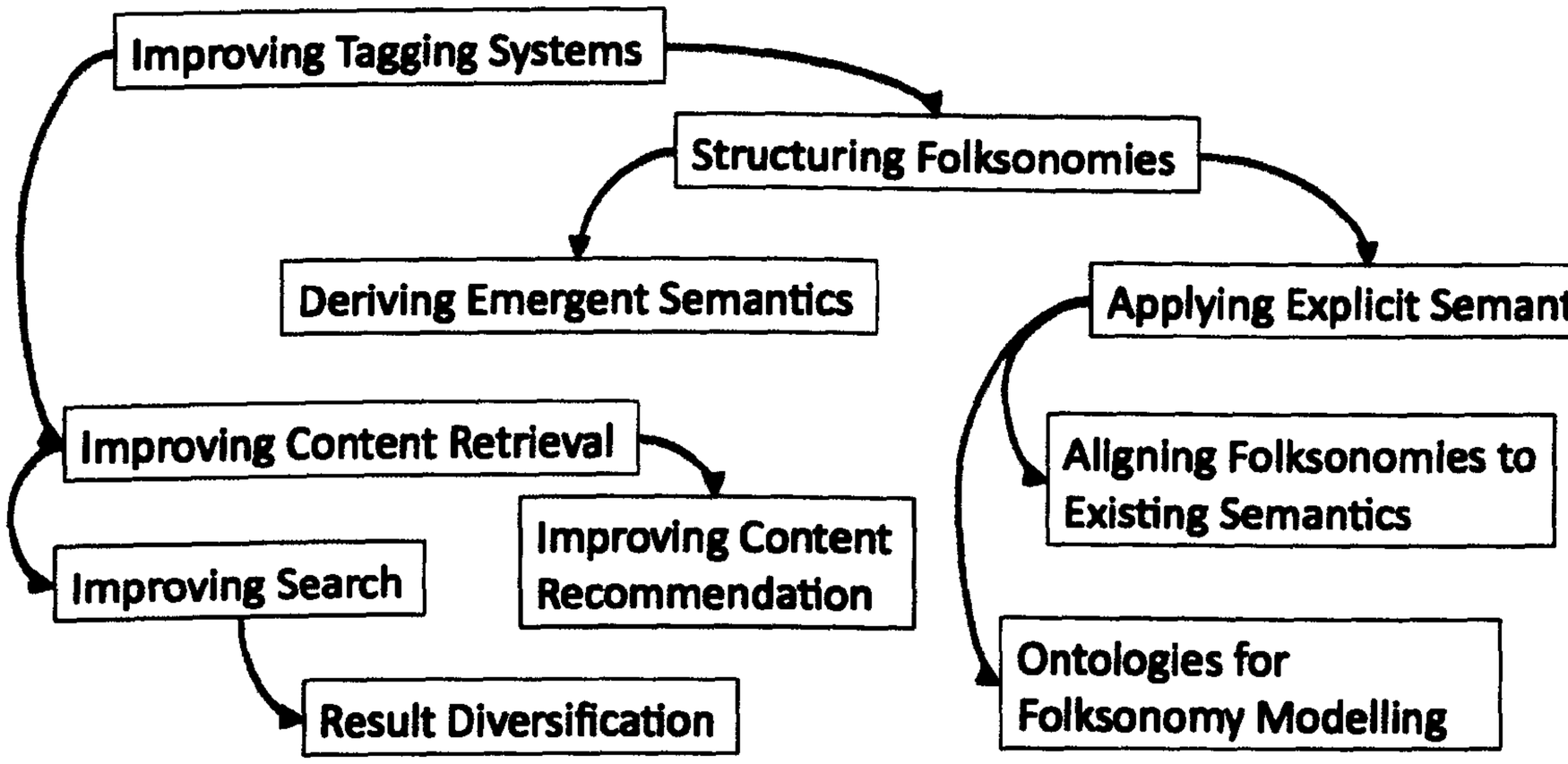


Figure 2.1: Overview of the literature on Improving Tagging Systems

to structuring folksonomies results in an interlinked graph, where the relations among the entities are named. Such structuring has been approached either by aligning folksonomies to existing structured resources (such as ontologies, thesauri or community maintained wikies) or by creating new models to represent them. In the following sections we present the most representative works on implicit and explicit folksonomy structuring.

Deriving Emergent Semantics From Folksonomies

The most vital characteristic of folksonomies, which allows for implicit semantics to emerge, is their network dynamics, such as the co-occurrence frequency of tags, resources and users in existing tagspaces. The hypothesis that "*entities which co-occur with a high frequency are somehow related*" lies beneath the notion of emergent semantics. A variety of techniques and measures have been applied for acquiring a better understanding of the emergent semantics. Such techniques are applied on folksonomies post-tagging, and as a result, they do not propose a shift in the tagging paradigm for the purposes of applying structure. A common characteristic of the works discussed in

this line is that, although hierarchical structures are discovered among the tags, their relations are not formally described.

One of the earliest studies, presented by Mika [86], represents the folksonomy network as a tripartite graph, whose nodes are the users, resources and tags and the edges are the connections among them. For example, an edge between a user and a tag node represents the usage of the tag from the user, while an edge between a resource and a tag node designates that the tag is used to tag the resource. Mika uses the tripartite graph to extract two graphs, a network of users based on commonly used resources and a lightweight ontology of tags(concepts) based on the overlap of the users and resources with which they co-occur.

Begelman et al. [26] also represent the tagspace as a weighted undirected graph, based on the tag co-occurrence frequency. They then apply spectral clustering to induce clusters of related tags. This method only results to related clusters but provides no indication on types of specific relations, such as hierarchical.

Schmitz [109] derives a hierarchy of tags by exploiting tag co-occurrence with a probabilistic model for subsumption generation. The basic hypothesis on which he bases his method is that a tag x subsumes a tag y if the probability of appearance of x given y is higher than a given threshold and the opposite is lower.

Cattuto et. al [36] perform an extensive analysis on different kinds of relatedness measures and ground the emergent relations by comparing them to the hierarchy of WordNet. They demonstrate that by using the appropriate measure different types of semantic relationships, such as synonymy and subsumption, can be obtained.

Halpin et. al [111] show that the distribution of tags follows the power low and stabilises over time. They use the stabilised tags and resources to extract hierarchies from folksonomies. They assume that the stabilised tags on a resource describe a general consensus on the topic of this resource. Using a number of heuristics they extract

hierarchical relations among tags based on their co-occurrence and their "information value", i.e., the number of resources tagged with them.

Heymann et. al [59] compose tag vectors based on the number of users who tagged each resource. Using cosine similarity to measure the distances among vectors they calculate betweenness centrality on the similarity graph of tags. Tags with higher centralities are added in the hierarchy first to represent more abstract concepts.

Zhou et. al [135] present an unsupervised method to extract hierarchical semantics from tags. They break down clusters of tags using deterministic annealing until they encounter a number of "effective clusters". These represent clusters whose semantics can be generalised by some tags, the "leading tags". For instance, in the cluster {`music`, `reggae`, `lyrics`}, music is the leading tag. They deduce hierarchical semantics using their notion of leading tags.

Wu et. al [124] represent the users, resources and tags as multidimensional vectors and place them in a multidimensional space of domains. They create links between domains and items according to the relations of tags, users and resources to these domains. The domains are identified through tag clustering which, along with the positioning of items within the domain space, is carried out dynamically.

The above works exploit tag statistics in a manner bound to introduce the "popularity vs generality" problem, where the popularity of a tag can be mistaken for generality and as a result induce a wrong hierarchy. Plangprasopchok et.al [102] use additional information to induce global hierarchies from personal user specified hierarchies. Using Flickr as their use case, they assume that the structure users apply when uploading and organising content into sets and collections generally reflects subclass and part-of relationships. Using graph and lexical similarities they merge the hierarchies of individual users and learn a global folksonomy. However, the authors highlight a key issue with their approach, the phenomenon that only a small percentage of user apply such organisations to their content. Hence data sparseness affects their results.

In another effort to create structures using additional information to user tags, Kim et. al [62, 65] propose a folksonomy contextualisation method based on Formal Concept Analysis aiming to provide shared meaning and create conceptual hierarchies from tags in the blogosphere. They base their work on the assumption that if a blog has relationships with others, they would use a similar set of tags. They deduce that contextualised folksonomies are able to provide context-centric and shared collections of tags to semantically-interlinked online communities.

Applying explicit Semantics to Folksonomies

In this section we discuss a line of work that involves the alignment of folksonomies elements to existing semantics or the creation of semantic models to represent folksonomies. The presentation of schemas for enriched annotation, a priori to tagging, propose a novel semantically-enabled way to tag and annotate. A disadvantage of structuring folksonomies with explicit semantics is that if the appropriate semantics does not exist (because of the lack of semantic entities describing the meaning of a novel tag, like "rss") human effort is required for their creation. An advantage of explicit over emergent semantics approaches is the potential to "publish" semantically described folksonomy content on the web of data according to the Linked Data principles [12].

Aligning Folksonomies to Existing Semantics

The assumption behind this line of work is that "*there exists an appropriate source of semantics for the representation of elements*" in folksonomies. One way to apply semantics to tagspaces is by **altering the paradigm of free tagging** and requiring the users to explicitly align their tags to some sort of semantic entity. Marchetti et.al [82] propose a semantic tagging system, SemKey, where users select a meaning for their tag from WordNet and Wikipedia. In the same line, Passant et. al [98] propose the "Meaning of a Tag" (MOAT), a framework that allows the users to link their tag to semantic entities from DBpedia or other ontological repositories. Passant

also [97] proposes a scheme for linking tags from blogs to ontologies. He presents an architecture that enables the authors of corporate blogs to select entities that most accurately represent the meaning of each tag for the particular blog post. The entities are provided from a pool of preselected domain ontologies. Along the same lines, Limpens et.al [75] propose an RDFS model to formalise the meaning of tags exploiting user feedback.

The following works also propose an alternative tagging process by requiring the user to provide more descriptive tagging. They then use the richer descriptions to create their own hierarchies which serve as semantic structures for their systems. Quintarelli et.al [103] describe FaceTag, a faceted annotation framework where users are encouraged to import hierarchies of tags, which are then used to improve search and various other functionalities. Along the same lines, Yoo et. al [131] implement a framework that allows the users to input more specific and more generic tags rather than a simple tag. Exploiting this information they provide a set of rules based on which, the aforementioned entries are turned into a hierarchy.

A less user intensive manner to apply semantics on the tagspace is **automatically aligning tags to semantic resources** and is usually applied to tagging spaces a posteriori to tagging. Maala et. al [79] distinguish six conceptual categories of tags. Using WordNet and other knowledge resources representing a category they organise the tags accordingly. Then, they enrich the Flickr photos with RDF triples specified for each of the conceptual categories. These triples are generated either by predefined predicates or from WordNet signatures depending on the category.

T-ORG [16] is a tag organisation system proposed by Abbasi et.al. Using a predefined ontology they categorise the tags, and as a result the resources they annotate, under the most appropriate ontological entity. They extract the ontological concepts and look for semantic relatedness between these concepts and the tags by combining them into predefined linguistic patterns and querying the web. Then each tag, and as a result

each resource, is categorised under a superclass of the concept to which it was most related by the web search.

Laniado et.al [69] transform the related tags cluster of a tag as provided by Delicious into navigable hierarchical structures. Using a combination of WordNet based metrics they identify the most appropriate synset for each tag. Then they extract the path of this tag from the WordNet hierarchy and they integrate it into the semantic tree which is built for each cluster.

Szomszor et al. [115] use a broader semantic source to model user interests based on the tags he has used. They align them to Wikipedia articles and use FOAF [9] notation to integrate and publish them.

Finally a hybrid approach that aligns tags to semantics after statistically processing them is presented by Specia and Motta [76]. They propose a method for extracting ontologies from folksonomies. Specifically they derive clusters of related tags and then align the tags to ontologies in order to discover their formal relations. Although not fully automated, this approach returns named relations between the semantically related tags in the clusters, thus formally adding semantics to the resulting structure.

Ontologies for Folksonomy Modelling

A complementary approach to aligning existing semantic entities to folksonomy elements is the creation of formal models to represent folksonomic elements, as well as their inter-relations. As early as the spreading of folksonomies, Tom Gruber [54] presented a formal model of elements and activities in folksonomies. In the TagOntology he proposes the expression:

Tagging (Object, Tag, Tagger, Source)

as a formal representation of the tagging activity. The meaning of the above is that, the tagger (user) tags the object (resource) with the tag in the context of the source system

(e.g., Flickr, Delicious etc.). The same year (2005) Newman [90] proposed an ontology which models folksonomy elements in the same manner with Gruber's approach, but also allows for the encoding of relationships among tags, such as "relatedTag" and "skos:broader".

Following Newman's example, Passant presented "Meaning of a Tag" MOAT [98] and Kim et.al the "Social Semantic Cloud of Tags", SCOT [64], which also represent the tagging activity from the perspective of a user (local) but also from the viewpoint of a community (global). These models also allow for explicit tag-to-meaning connection. Kim, Passant et.al [63] present an alignment of the most popular tag ontologies and demonstrate how this can support folksonomy modelling but also content reuse in various use cases across web applications.

The latest advance in the ontological modelling of tagging has been proposed by the CommonTag initiative [3]. This approach differs from the previous by being more straightforward from an implementation perspective. It proposes the embedding of semantic metadata into the HTML code of a webpage using RDFa. It presents a richer representation for tags that allows the tagging of a whole document or components of it by embedding semantic metadata into the tags. The meaning of the tag is decided during the webpage creation by assigning to it DBpedia or other entities that carry semantics. The CommonTag initiative is supported by large web companies such as Yahoo and Freebase.

2.3.2 Improving Content Retrieval in Folksonomies

The majority of work investigating folksonomies was motivated by Golder and Huberman who distinguished the three main issues that affect content retrieval, polysemy, synonymy and basic level variation (see Chapter 3 for the detailed explanation of these phenomena). To overcome these issues and improve search, the research community has focused its efforts on two research lines. The first line relies on **statistical meth-**

ods exploiting the dynamics of folksonomies and the distribution of tags, resources and users in order to address the problems in search caused by the above phenomena.

Bender et.al [27] present a hyper-graph of folksonomy entities and define their weighted inter and intra relations to entities of the same and the other types (e.g., user to content, user to user, user to tag, tag to tag and so on). They calculate the weights using entity co-occurrence. For each user query, they perform six types of query expansion based on the similarities of entities calculated from the hyper-graph and return the documents mostly relevant to the query and the user [3]. Experimenting with data from Delicious and Flickr on Precision @ 10 they show that the improvement is significant when using expansion based on the social network, i.e. seeking resources from the user's friends who have used the same tags as his query. De Meo et. al [85] also exploit the tripartite structure of folksonomies to create tag-resource and tag-user graphs which they use to perform query expansion. This method requires double relevance feedback on the user query. Upon submitting the query the most authoritative tags are selected based on the above graphs. The user needs to select which are the most relevant. These are further used to retrieve resources whose relevance is once again judged by the user. The intensive relevance feedback is required for enriching user profiles which can also support resource recommendation. Zanardi et.al [133] propose an approach for query expansion using collaborative filtering. In their approach a similarity measure is defined for the tags based on their frequency on resource labelling, Social Ranking. Each user query is expanded using the most similar tags to the query keywords based on K-Nearest Neighbour algorithm. Abbasi et. al [17] use a modified version of the vector space model which aims at addressing the tag sparseness in Flickr and improve the search results especially for poorly annotated images. They associate the resources with relevant tags based on contextual and distributional similarity of their tags with the new ones. In this way, they enrich the annotation of the images rather than perform query expansion. Yet, their approach in discovering sparsely annotated images performs quite well.

[3]The same query from a different user yields different results

The second research line in folksonomy improvement aims to address the common underlying cause of polysemy, synonymy and basic level variation, which is the lack of **machine understandable descriptions for the meaning of the tags and their relations**. An alternative paradigm to statistically expanding user queries is proposed, which exploits existing knowledge sources. In particular the TagPlus system [73] uses WordNet to disambiguate the senses of Flickr tags by performing two step queries. First each query tag is matched against WordNet synsets which then are presented to the users. The latter select which senses they are interested in. The results consist of resources tagged with the synonyms of the selected by the user sense. In that way the problems caused by polysemy and synonymy are ruled out. Similarly, the SynTag system [72] uses WordNet to improve search and requires the users to connect tags to senses in annotation time rather than query time. Finally, a more recent approach, which exploits ontologies rather than WordNet, is presented by Pan et. al [95] and aims at improving the precision of results, i.e., addressing polysemy. The queries are expanded with more descriptive terms extracted from a preselected ontology and the results are narrowed down to the resources that are tagged with the more descriptive set of terms. The association of query terms to ontological entities is performed automatically using tf/idf, and once the most appropriate entity is located one of the preselected query expansion strategies (Individual Property Value Expansion, Individual Class Expansion and Individual Property Expansion) takes place, in order to obtain the expanded query.

Additional functionalities have been proposed for improving the search experience in folksonomies. In an era of information overload classic IR improvements in terms of precision and recall should be complemented with enhancements such as **result diversification** [42]. Result diversification is the meaningful organisation of results, for example in cases of polysemy. Recognising this need Flickr provides a built[4] in result organisation for popular tag queries. The exact clustering algorithms used in Flickr are not known, yet the approach of van Zwol [121] achieves similar result diversification

[4]Accessed on 23/08/2010: `http://www.flickr.com/photos/tags/apple/clusters/`

using statistical measures on tags. An alternative approach presented by van Leuken et.al [119] is to use image clustering algorithms that exploit the visual similarity of images. In the latter the number of clusters depends on the differentiation on image characteristics and as a result it is less likely to rule out clusters representing a less popular sense for a tag (see the example for the query **orange** in Section 3.3).

Finally apart from search improvement methods, further work has focused on optimising the **recommendation of relevant content**. Content discovery via social connections is a useful functionality enabled by the social dimension of folksonomies [120] and there is significant research towards this direction. Amer-Yahia et al. [126] produce recommendations of resources, tags and users by calculating the similarity among users based on two measures; the number of tags and the number of resources they share. Tso-Sutter et al. [118] perform resource recommendation based on collaborative filtering of resources. They extend traditional collaborative filtering using shared tags among the users. Firan et al. [47] and Zhao et al. [134] produce recommendations based on user interest similarities as follows. They use a WordNet based similarity metric of the tags used by two users in order to deduce similarity among their interests. Carmagnola et al. [35] also use WordNet to classify tags in various categories (subjective or novel tags) and use these tags to asses user behaviour based on the profiles built with them.

2.4 Summary and Outlook

In this section we give a summary of the existing approaches focusing on the methods they used. In addition, we outline the approaches for evaluating the proposed techniques and conclude by describing the open issues in the literature.

2.4.1 Employed Methods

In Sections 2.2 and 2.3 we described the most significant research conducted in the scope of understanding and improving tagging systems. In this section we present an overview the most popular methods employed for their study. For the sake of consistency with the goals of this thesis we separate them in methods using statistics and methods using semantics.

The majority of work presented in Sections 2.2 and 2.3 largely employ **statistical measures**. One of the reasons such methods are widely required is the scale of data generated from tagging systems. Statistical methods allow for macroscopic analysis of tagging systems and yield a holistic and high level understanding. One of the statistically enabled approaches represents folksonomies as graphs. Usually, the nodes of the graph represent the users, tags and resources while the edges represent their interactions and interrelations. Most of the time the edges of the graphs are weighted based on the frequency of certain interactions, such as the frequency a user has used a tag, the number of tags she assigned to a resource, the number of users that use a specific tag to annotate a resource and so on. Frequency of interactions is also used in vector space models and latent semantic analysis where, instead of using graphs, the folksonomy elements (mainly the tags) are contextualised using their frequent co-occurrence in different contexts. With regards to tag contexts, the majority of work uses either the resource-based, which is the tags co-annotating resources with the tag of interest, or the user-based, which comprises of the tags the user has used along the tag of interest. Various combinations of the two have also been used and improved with the social context of a user, which consists of tags used by her network. Additional probabilistic approaches also exploiting co-occurrence have been proposed. Finally, other approaches exist that do not solely exploit co-occurrence of entities but employ statistical measures to perform image and text analysis. As an attempt to generalise, we observe that statistical methods were extensively used to extract emergent semantics, perform query expansion, search diversification and content recommendation in

folksonomies.

On the contrary, considerably less work has used **explicit semantics** to study the tagging systems. On the one hand the semantically-enabled methods for the web are less mature compared to statistical analysis. On the other hand the microscopic study and semantic description of each folksonomy element is less efficient towards understanding the tagging systems and extracting their implicit semantics if we take into consideration their large scale.

The motivation for most of the work focused on extracting semantics from tagging systems is to exploit data derived from folksonomies in order to extract semantics for the purposes of improving existing web functionalities (such as web search) and learning hierarchies, rather than improve the functionalities of folksonomies (such as search, content organisation, annotation and so on). As a result, the approaches that use explicit semantics are limited to aligning tags to WordNet and ontologies or creating new ontologies to represent the interactions within tagspaces. However, the collective element behind Web2.0 has allowed for the creation of resources, such as Wikipedia, DBpedia and Freebase. These, although they are less formal and often are characterised by their lightweight semantics, have also been used to study and improve folksonomies.

2.4.2 Evaluation Approaches

In contrast to other research areas where evaluation benchmarks have been established (Multilingual Information Retrieval [5] or Ontology Alignment [13]) and widely adopted, the folksonomy research field is lacking such common initiatives mainly due to its young age. Hence, each work is evaluated with a different strategy and dataset, and independently of the rest. For the same reason, a small percentage of the approaches has been extensively evaluated while the majority provide qualitative or empirical results.

The work focused on improving search in folksonomies employs the **classic IR measure** of precision to evaluate the results. Instead of measuring the recall, due of the lack of the ground truth, which is virtually impossible to obtain in social networks, these approaches employ the notion of rate of increase of relevant results. These are obtained by comparing a baseline-type of search (such as the search in a folksonomy system) to the search enabled by the method they propose [17]. To validate such methods **user-based studies** are conducted where humans are required to provide relevance judgements [22, 27] .

In terms of evaluating structured folksonomies, to the best of our knowledge, the only study which presents an **evaluation of the learned hierarchical structure** is presented by Plangprasopchok et. al in [102]. They validate their learned folksonomies against the Open Directory Project[5] hierarchy using Lexical Recall [80] and a modified version of Taxonomic Overlap [101].

With the maturing of the field, and the increasing number of relevant approaches, the creation of a common evaluation framework would be highly beneficial despite the complexity of such task due to the heterogeneity and multidimensionality of tagging systems. However, its existence would largely benefit the related research, minimise repetitions and enable more rapid progress.

2.4.3 Open Issues

Overall, a wide variety of approaches aims to alleviate the limitations of free tagging of folksonomies. Statistic and semantic approaches are useful and allow for diverse ways of improvement in tagging systems. However, some issues still remain open.

- Approaches that use the social dimension of tagging systems have proved to be very successful in tag disambiguation and search improvement. However, such

[5] `http://www.dmoz.org/`

approaches depend not only on the explicit social relations (a user belongs to the list of friends of another), but mainly on the implicit relations emerging via shared tags and resources. Such rich social structure is only available in broad folksonomies [78] such as Delicious, Last.fm and Citeulike, where one resource is usually annotated by more than one user. In cases of narrow folksonomies, such as Flickr and Youtube, these approaches do not apply. Therefore there is a need for **approaches that are independent of social structures.**

- Approaches that structure folksonomies with explicit semantics require some sort of human effort. A line of research, proposes a paradigm shift in free content annotation by the user [73, 103], which cancels the unrestricted nature of free tagging and requires more effort from the side of the user. The approaches that automatically align tags to existing semantic structures require some type of initialisation [16], i.e., selection of the appropriate ontologies that represent a relevant domain to the tagspaces. As a result, **methods that automatically select the appropriate knowledge** are needed.

- The statistically derived structures from folksonomies are frequently useful only within a particular folksonomy ecosystem and can only be exploited by a tailored technique [17, 59, 109]. On the contrary the **structuring of folksonomies using explicit semantics allows for a persistent description of their elements**. In this way, the latter may lend themselves to cross domain, cross system applications and can be potentially exploited by endless semantic applications.

As a result we identify the need for an approach, that automatically applies structure on tagging systems, is uninfluenced by social parameters, and is domain independent.

As Hendler and Golbeck [58] point out:

"Some techniques have tried to add structure to tags using clustering methods. Though this can sometimes create sensical "hierarchies", the links between concepts do not

indicate parenthood as we would normally expect (...) This kind of hierarchy will not significantly improve search and information structure as well as one that is human engineered. The first challenge, then, is how to build a structure around tags (...) that goes beyond clustering methods ".

In the next chapter we present the formalisation of the concepts employed in our thesis.

Chapter 3

Problem Formalisation and Definitions

In this chapter we provide definitions for the folksonomic entities that are involved in our research and their relations. We describe their characteristics and the issues that influence folksonomy search. We introduce the key concepts of our approach and propose measures for the evaluation of sense structures and the performance of the enrichment algorithms. Finally, we present an ontology which is used to describe the output of the enrichment process.

3.1 Introduction

In this chapter we give a detailed analysis of the problem this thesis addresses. We present the main entities of folksonomies and their dependencies (Section 3.2). We introduce the main issues that emerge from the paradigm of free tagging and describe the limitations they pose in folksonomy search (Section 3.3). We introduce the concepts and entities used in this thesis for the semantic enrichment of folksonomies (Section 3.4) and specify an ontology that represents the semantically enriched tagspaces and the enrichment algorithms (Section 3.5). Finally we present measures for the evaluation of

the semantically enriched tagspaces (Section 3.6). To illustrate our definitions we use Figure 3.1 throughout this chapter as a running example. The left hand side represents a folksonomy tagspace and the right hand side a semantic structure that defines the meaning of each tag and the relations among the senses.

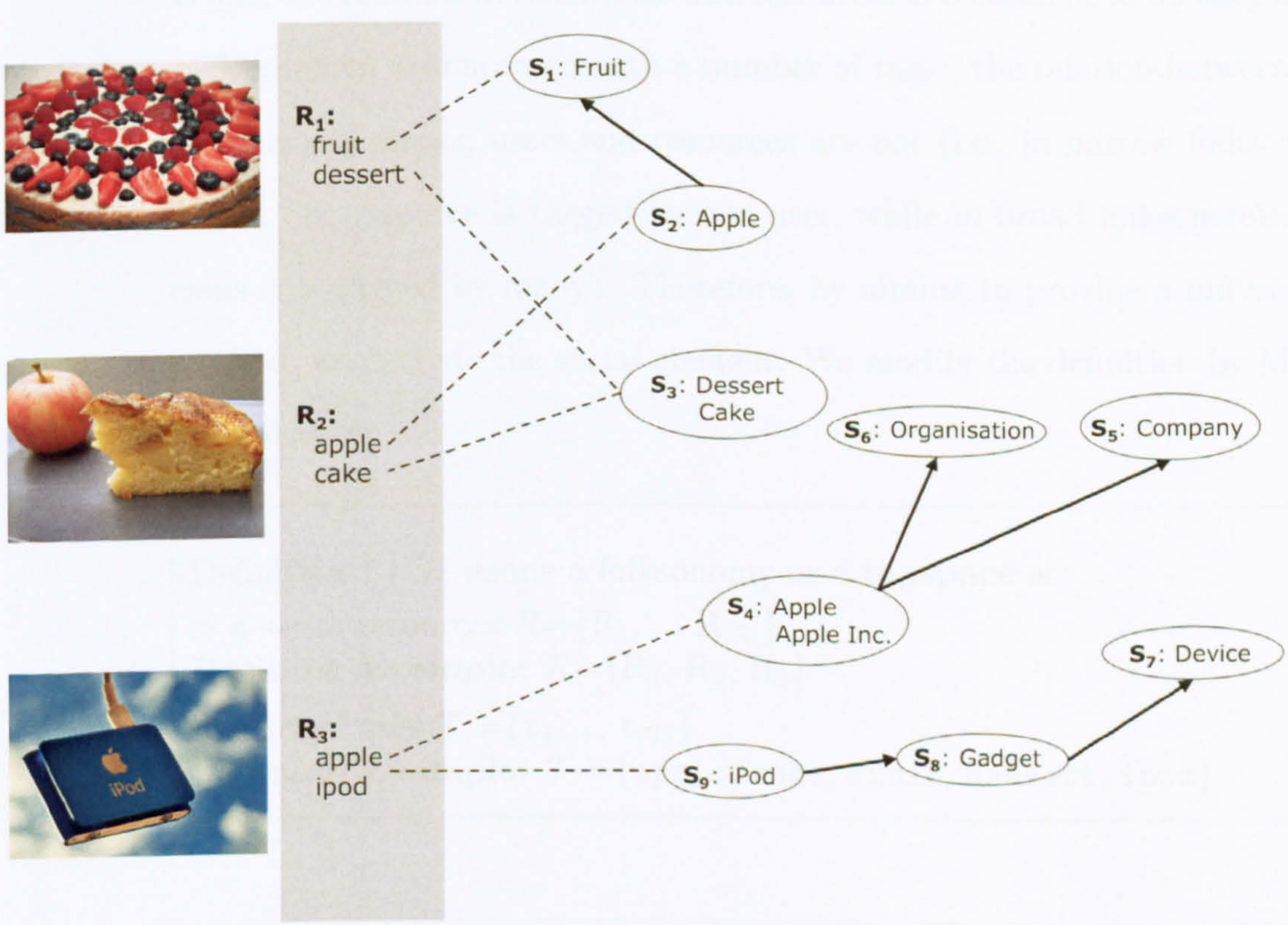


Figure 3.1: Example folksonomy with three resources and a structure of senses representing the meaning of the tags and their relations

3.2 Folksonomies

The elementary entities of folksonomies are the resources, the users and the tags. The content of the **resources** depends on the folksonomy of reference and can be images [8], audio tracks [11], video clips [15], bookmarks [7] and more. The **users** provide annotations for the resources during the **tagging** process. The annotations consist of one or more **tags**, which are words freely chosen by the users to describe the resources according to their own perception. Mika [86] provided the first formal definition of the

tripartite model of folksonomies, which was further adopted and extended by the community[1]. His model includes all three elements of folksonomies, users, resources and tags however, our approach is user-independent. In particular, our approach is resource and tag centric and does not take into account the social dimension of folksonomies. While, the relation between tags and resources is consistent in all tagging systems (i.e., each resource is annotated with a number of tags), the relation between users and tags and the one among users and resources are not (i.e., in narrow folksonomies, such as Flickr, one resource is tagged by one user, while in broad folksonomies, such as Delicious it is tagged by many). Therefore, by aiming to provide a universally applicable method, we exclude the social element. We modify the definition by Mika and specify the following.

Definition 1. We define a folksonomy or a **tagspace** as:
▷ a set of resources $\mathcal{R}=\{R_1, .., R_{|\mathcal{R}|}\}$.
Running Example: $\mathcal{R}=\{R_1, R_2, R_3\}$
▷ a set of tags $\mathcal{T}=\{t_1, .., t_{|\mathcal{T}|}\}$.
Running Example: $\mathcal{T}=\{\texttt{apple}, \texttt{cake}, \texttt{fruit}, \texttt{dessert}, \texttt{ipod}\}$

Definition 2. We represent the **relations among tags** and **resources** using the following functions:
▷ *tags*(R) represents the set of tags associated with resource R.
Running Example: *tags*(R_2)={`apple`, `cake`}

▷ *res*(`t`) is the set of resources associated with tag `t`. $1 \leq |res(\texttt{t})| \leq |\mathcal{R}|$; each `t` $\in \mathcal{T}$ can annotate at least one and at most all the resources in the folksonomy. For the sake of clarity we adopt two naming conventions for the tags:

- We characterise as **specific** those tags `t` that belong to a resource R and we annotate them with the resource name, i.e., `R_t`.
- If we do not refer to the instance of a tag that is assigned to a specific resource we characterise it as **generic**, i.e., `t`

Running Example: *res*(`apple`)={R_2, R_3} and *res*(`R2_apple`) = {R_2}.

[1] *Ontologies Are Us: A Unified Model of Social Networks and Semantics* 555 citations on August 2010

Figure 3.2: Clustered results from Flickr when querying with `apple`

We represent the co-occurrence of a tag with other tags using the notion of **context**.

Definition 3. We represent the **context of a tag** as:

▷ resource-based context: The set of tags assigned to the same resource as `R_t`. The resource-based context is also denoted by the **tagset** of R which is $T_R = tags(R)$.
Running Example: $T_{R_2} = tags(R_2)$={`apple`, `cake`}

▷ cluster-based context. The **cluster**, C, is a more generic set of tags that has emerged from the statistical analysis of the tagspace $\mathcal{T}$. A cluster of tags is created based on their global co-occurrence frequency with resources or users. A cluster of tags is rarely associated with one resource, it is frequently associated with a set of highly similar resources, from whose annotations it usually emerges.

Figure 3.2 depicts the clustered results returned from Flickr for the query `apple`. We note that each group of images is associated with a cluster of tags on the right hand side. Each tag cluster contextualises the tag, e.g., `apple` in a different dimension. In fact, each of these clusters contextualises `apple`, while the tagset of R_2 contextualises `R_apple`. The last row of results in Figure 3.2 contains cluster C = {`nyc`, `newyork`,

`manhattan`, `newyorkcity`, `ny`}, which is the cluster of tags co-occurring with `apple` and implies the sense of *Big Apple* (New York).

3.3 Main Issues in Folksonomies

One of the major assets of folksonomies, which is the free tagging, is also the cause of the three phenomena that largely impact on content organisation and retrieval. Initially highlighted by Golder and Huberman [51], these are:

Tag synonymy arises when lexically different tags express the same concept. For example `cake` and `dessert` usually express a sweet meal course. Synonymy may cause exclusion of results if these are tagged with synonym(s) of the search keyword. For instance, searching for `cake` in the example of Figure 3.1, only returns R_2 and not R_1 even if this is correct match.

Tag polysemy occurs among lexically identical tags that denote different meanings. For example, `apple` may refer to fruit or a brand name. Tag polysemy causes the retrieval of unwanted results when the tag is used with a different meaning than the search keyword. Searching for `apple` in the folksonomy of Figure 3.1 will return both R_2 and R_3, although, depending on the meaning of `apple` in the context of the query, only one of these resources is relevant.

Basic level variation. Tags with different levels of specificity are used to describe resources that relate to the same concept. For example, `apple` and `fruit` can both describe resources about apples. The lack of structure in the tagspaces, does not allow for explicit declaration of the fact that "*apple is a fruit*". This limits the potential of querying for resources tagged with related tags. For the example of Figure 3.1, querying for `fruit` only returns R_1, although R_2 is also tagged with a fruit name.

Figure 3.3: The most interesting results for the query `lake europe` in Flickr

However, the problems of content retrieval in folksonomies are not limited to the above examples. Consider a Flickr user who wishes to search for pictures of lakes in europe. In lack of more elaborate query mechanisms, her only option is to use the keyword search with arguments {`lake europe`}. The most interesting[2] results[3] for this query are presented in Figure 3.3. The system is not able to provide a meaningful categorisation of the results or some kind of recommendation such as {`lake austria`}, {`lake greece`}, {`lake balkans`} because the european countries or areas that contain lakes are not associated with the concept of europe.

Some existing folksonomy search systems are able to provide a more meaningful categorisation and diversification of the results, as shown in Figure 3.2, which displays the results of the query `apple`. We note that there is one cluster for the sense of *Fruit*, one cluster for *Big Apple* and two clusters for *Apple Inc.*. This is because of the higher co-occurrence of the tag `apple` with *Apple Inc.* related terms. The two clusters could depict narrower senses of *Apple Inc.* which are the two main product lines, computers

[2] `http://www.flickr.com/explore/interesting/`
[3] This query returned 42.710 results (September 2010)

Figure 3.4: Clustered results from Flickr when querying with `orange`

and hand-held devices. However, this approach is biased by the popularity of tags in the tagspace. In other words, if one of the meanings of a polysemous tag does not occur frequently in the tagspace then it is eliminated. This is apparent in the search results for `orange`, depicted in Figure 3.4. There, the sense of *Colour* occupies all four clusters of results, while the sense of *Fruit* is not represented. This is because the *Colour* sense of the tag is disproportionally frequent to the *Fruit* sense.

Enriching tagspaces with a semantic structures that make the meaning and relations of tags explicit, in combination with appropriate query mechanisms can help alleviate such problems. In the next section we introduce the concepts related to the semantic enrichment that we employ throughout this thesis.

- S: (n) **apple** (fruit with red or yellow or green skin and sweet to tart crisp whitish flesh)
  - *direct hyponym* / *full hyponym*
    - S: (n) crab apple, crabapple (small sour apple; suitable for preserving) *"crabapples make a tangy jelly"*
    - S: (n) eating apple, dessert apple (an apple used primarily for eating raw without cooking)
    - S: (n) cooking apple (an apple used primarily in cooking for pies and applesauce etc)
  - *direct hypernym* / *inherited hypernym* / *sister term*
    - S: (n) edible fruit (edible reproductive body of a seed plant especially one having sweet flesh)
    - S: (n) pome, false fruit (a fleshy fruit (apple or pear or related fruits) having seed chambers and an outer fleshy part)
  - ***part holonym***
    - S: (n) **apple**, orchard apple tree, Malus pumila (native Eurasian tree widely cultivated in many varieties for its firm rounded edible fruits)

Figure 3.5: A WordNet synset for *Apple*.

3.4 Semantically Enriching Folksonomies

In this thesis we address the underlying cause for all the aforementioned folksonomies issues which is the **lack of structure**. In contrast to previous work, we reuse all the available knowledge encoded in Knowledge Sources and automatically apply semantics to tagspaces. In the following we define the core concepts of this work.

Definition 4. We define as **Knowledge Source**:
▷ a body of knowledge that contains semantic descriptions of concepts and explicitly defined semantic relations between them.

One such well established Knowledge Source is WordNet, which has been widely adopted in various research areas, including folksonomy improvement (see Chapter 2). In addition to WordNet, ontologies, are hand-crafted artefacts that also provide explicitly formalised knowledge. In the scope of this thesis we utilise WordNet and ontologies publicly available on the web through the Watson Semantic Web Gateway [41].

Definition 5. We define **Semantic Entity** as:
▷ a Knowledge Source object that contains information that defines a concept.

Details for http://ontosem.org/#apple

- In http://morpheus.cs.umbc.edu/aks1/ontosem.owl
  - *Class*
  - **label**: "the firm, rounded fruit of a tree, having skin that is usually red but may be yellow or green"
  - **subClassOf**: http://ontosem.org/#tree-fruit

Figure 3.6: A Class derived from the Ontosem.owl ontology for *Apple*.

An example of a semantic entity is depicted in Figure 3.5[4] in the form of a WordNet noun synset. Ontological Classes and Individuals are also semantic entities. An example of ontological semantic entity is depicted in Figure 3.6 [5]. In both cases, the semantic entities describe the concept of *Apple* along with semantic relations to other concepts. For example, in Figure 3.5 *Apple* is a hyponym of *Edible Fruit* and *Pome*, while in Figure 3.6 it is a subclass of *Tree Fruit*. In the rest of the thesis we use the following notations to represent synsets and ontological entities respectively.

Synset: ***apple*** $\overset{hyponym}{\rightarrow}$ ***edible fruit*** $\overset{hyponym}{\rightarrow}$ ***fruit*** $\rightarrow$ **..**
$\overset{hyponym}{\rightarrow}$ ***pome*** $\overset{hyponym}{\rightarrow}$ **..**
$\overset{hypernym}{\rightarrow}$ ***crabapple***
$\overset{hypernym}{\rightarrow}$ **...**
"fruit with red or yellow or green skin [...] tart crisp whitish flesh."

Class: ***apple*** $\overset{subClassOf}{\longrightarrow}$ ***tree-fruit*** $\overset{subClassOf}{\longrightarrow}$ **...**
(Ontology 1) *"The firm rounded, fruit of a tree [...] may be yellow or green."*

The related concepts of a semantic entity constitute its **semantic neighbourhood**. In the following chapters we describe the extraction and usage of semantic entities for the definition of the meaning of tags.

In order to unify the heterogeneous semantic entities that originate from different

[4] http://wordnetweb.princeton.edu/

[5] http://watson.kmi.open.ac.uk/WatsonWUI/entity_look_up.html?q=http://ontosem.org/#apple

Knowledge Sources and integrate them into a coherent structure we define a new concept.

Definition 6. We define **Sense** as:
▷ an object that holds the meaning of a word. Each sense S has a number of **subordinate (subsenses - $sub(\boldsymbol{S})$)** and **superordinate (supersenses - $sup(\boldsymbol{S})$)** senses. In addition, it has a set of synonym terms $syn(S)$ that define its meaning.
Running Example: $sub(S_1)$={S_2}, $sup(S_4)$={S_5, S_6} and $syn(S_3)$ = {`cake`, `dessert`}

The two semantic entities of *Apple* described previously, are translated into senses in the following manner.

Sense: ***apple*** $\overset{hyponym}{\rightarrow}$ ***edible fruit***
W(1) $\overset{hyponym}{\rightarrow}$ ***pome***
$\overset{hypernym}{\rightarrow}$ ***...***
"fruit with red or yellow or green skin [...] tart crisp whitish flesh."

Sense: ***apple*** $\overset{subClassOf}{\longrightarrow}$ ***tree-fruit***
O(1) *"The firm, rounded fruit of a tree [...] may be yellow or green."*

The codes O(1) and W(1) represent the **provenance** of the sense. For example, the first sense was created by a WordNet synonym, while the second by an ontological class. In later chapters where we describe the sense integration, these indicators may co-exist in the same sense (i.e., the respective sense is created using three ontological entities and a WordNet synset). In Chapter 7 we describe how the relations of the newly created senses to their neighbour semantic entities are transformed to relations between senses, and subsequently integrated in a sense structure.

A sense can have zero or more supersenses or subsenses in a specific hierarchy. For example, there are no supersenses for S_1:*Fruit* or subsenses for S_2:*Apple*, hence $sup(S_1)$ = {} and $sub(S_2)$ = {}. Each sense may carry additional relations to other senses apart from supersenses and subsenses. These are further analysed in Chapter 7. Finally, the

actual implementation of a sense contains more information than depicted above. For example, it contains a set of de-reference-able URIs of the semantic entities from which it originates. We do not demonstrate this data in the above representations in order to avoid the visual clutter. More in-depth analysis of the senses is given in Chapter 7 where the process of sense creation is explained in more detail. Also, the description of the properties of a sense is presented in Section 3.5, where we introduce the ontology that supports the semantically enriched tagspaces.

Our approach associates tags to the correct meaning, i.e., sense. Two types of **relations can hold between a tag t and a sense S**[6] depending on whether the tag is specific or generic (see Definition 2). For the example of Figure 3.1, the tag `apple` can be defined either by S_2 or by S_4. Both S_2 and S_4 are **candidate senses** for `apple`, therefore:

$$\texttt{apple} \xrightarrow{hasPotentialDefinition} S_2 \text{ and } \texttt{apple} \xrightarrow{hasPotentialDefinition} S_4$$

On the contrary, when the tag is explicitly specified in the context of a resource tagset, such as $\texttt{R}_2\texttt{\_apple}$ and $\texttt{R}_3\texttt{\_apple}$, it can be related with at most one sense:

$$\texttt{R}_2\texttt{\_apple} \xrightarrow{hasDefinition} S_2 \text{ and } \texttt{R}_3\texttt{\_apple} \xrightarrow{hasDefinition} S_4$$

Definition 7. We define the **relationship between senses and tags** as definition (Dfn) of `t` from S:

▷ $Dfn(\texttt{t}, S)$, s boolean function which, for sense S and tag `t` is defined as:

$$Dfn(\texttt{t}, S) = \begin{cases} 1 & \text{if } S \text{ holds a meaning for } \texttt{t}, \\ 0 & \text{if not.} \end{cases} \tag{3.1}$$

Running Example: $Dfn(\texttt{apple}, S_2) = 1$ and $Dfn(\texttt{R}_3\texttt{\_apple}, S_2) = 0$

▷ $senses(\texttt{t})$ returns a set of senses (candidate or explicit) for tag `t` and $0 \leq |senses(\texttt{t})| \leq |\mathbb{N}|$

Running Example: $senses(\texttt{apple}) = \{S_2, S_4\}$ and $senses(\texttt{R}_3\texttt{\_apple}) = \{S_4\}$

[6]The notation `tag` and *Sense* has been used to reflect the fact that tags are plain text while senses carry semantics.

The empty set, $senses(\mathtt{t}) = \{\}$ is obtained if the tag is not enriched with any sense. It should be pointed out that $senses(\mathtt{R\_t}) = S$ does not imply $Dfn(\mathtt{R\_t}, S) = 1$. Even if there is only one candidate sense for `t`, it is not guaranteed that it is the correct one without considering `t`'s context.

The explicit or semantic relations between tags and senses (or entities) e are related to the notion of **coverage** of `t` by e. We assume that a sense or semantic entity e covers a tag `t` in the following cases:

Lexical Coverage holds when `t` belongs to the set of lexical descriptions of e. If e is an ontological entity, then the lexical descriptions consist of the local name and the labels of e. If e is a WordNet synset or a sense then the lexical descriptions are its synonyms. `t` can be lexically covered by many entities (senses) which means that each of them represents a candidate meaning for `t`.

Semantic Coverage holds when the meaning of `t` is explicitly described by e. In order to decide if e semantically covers `t`, `t` needs to be contextualised in some manner (either in the scope of a resource tagset, or in the scope of a cluster, Definition 3). The semantic coverage of `t` from e is equivalent to $Dfn(\mathtt{t}, e) = 1$ (see M3.7)

To conclude, in this section we introduced the concepts of **Knowledge Source**, which provides the enrichment algorithm with **semantic entities**. The semantic entities are transformed to **senses**, which are used to describe the meaning of tags and their relations. Although previous work exploits additional metadata of the resources [101, 102], we focus our experiments on the tags of the resources rather than other lexical attributes such as title, description and notes. Tags offer higher precision and recall to the queries in comparison to all other lexical information of the resource [78]. In addition, the types of lexical information vary among different folksonomies, while tags are consistent.

In the following section we describe the FLOR-ontology that is used to represent the enriched tagspaces.

3.5 An Ontology for Enriched Tagspaces

In order to support the output of our enrichment algorithm (FLOR-2, Chapter 7), we built the **FLOR ontology** with which we define the schema which supports the relations among tags, resources, senses and semantic entities as described in the previous sections. We create a new ontology for the representation of the enriched tagspaces, rather than reusing the ontologies described in Section 2.3.1. This is because these ontologies align the tag to a semantic entity directly and do not support the notion of sense. As we described in the previous section, the notion of sense is important for the integration of entities originating from different backgrounds and the creation of an interconnected semantic layer for the input tagspaces [7]. The enrichment algorithm described in Chapter 7 (FLOR-2) directly populates the ontology with RDF representations of the enriched tagspaces. Figure 3.7 is a visualisation of the main ontological classes and their relations. Each class is used to represent a set of fundamental entities involved in the enrichment process as described in the previous sections on this chapter. We specify each ontological entity in Appendix B. To help us exemplify the descriptions of the entities, in Figure 3.8 we present a more detailed and realistic representation of the semantic structure created by FLOR-2.

Resource$_X$ is tagged with {`X_Hungary`, `X_Balaton`, ... , `X_Europe`}. The tags of Resource$_X$ are matched to senses that define their meaning. Each of these senses relates with the senses of the tags in the same tagset for example:

$$\texttt{X\_Balaton} \xrightarrow{hasDefinition} Balaton \xrightarrow{partOf} Hungary \xleftarrow{hasDefinition} \texttt{X\_Hungary}$$

[7] The concept of "Meaning" from MOAT ontology is the closest representation to the concept of sense but does not provide representation of lexical forms and relations which are further required for the creation of semantic layers

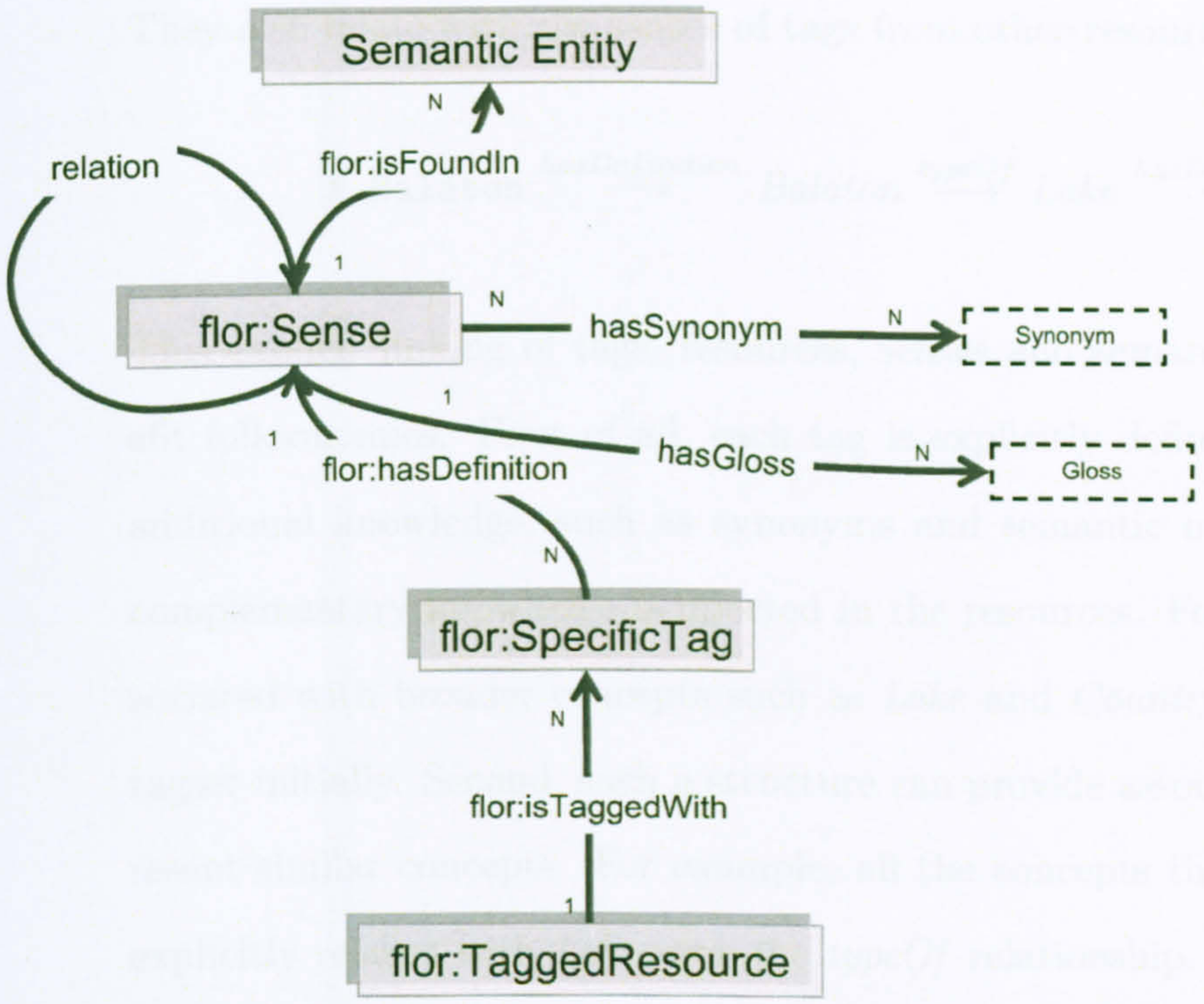


Figure 3.7: FLOR Ontology

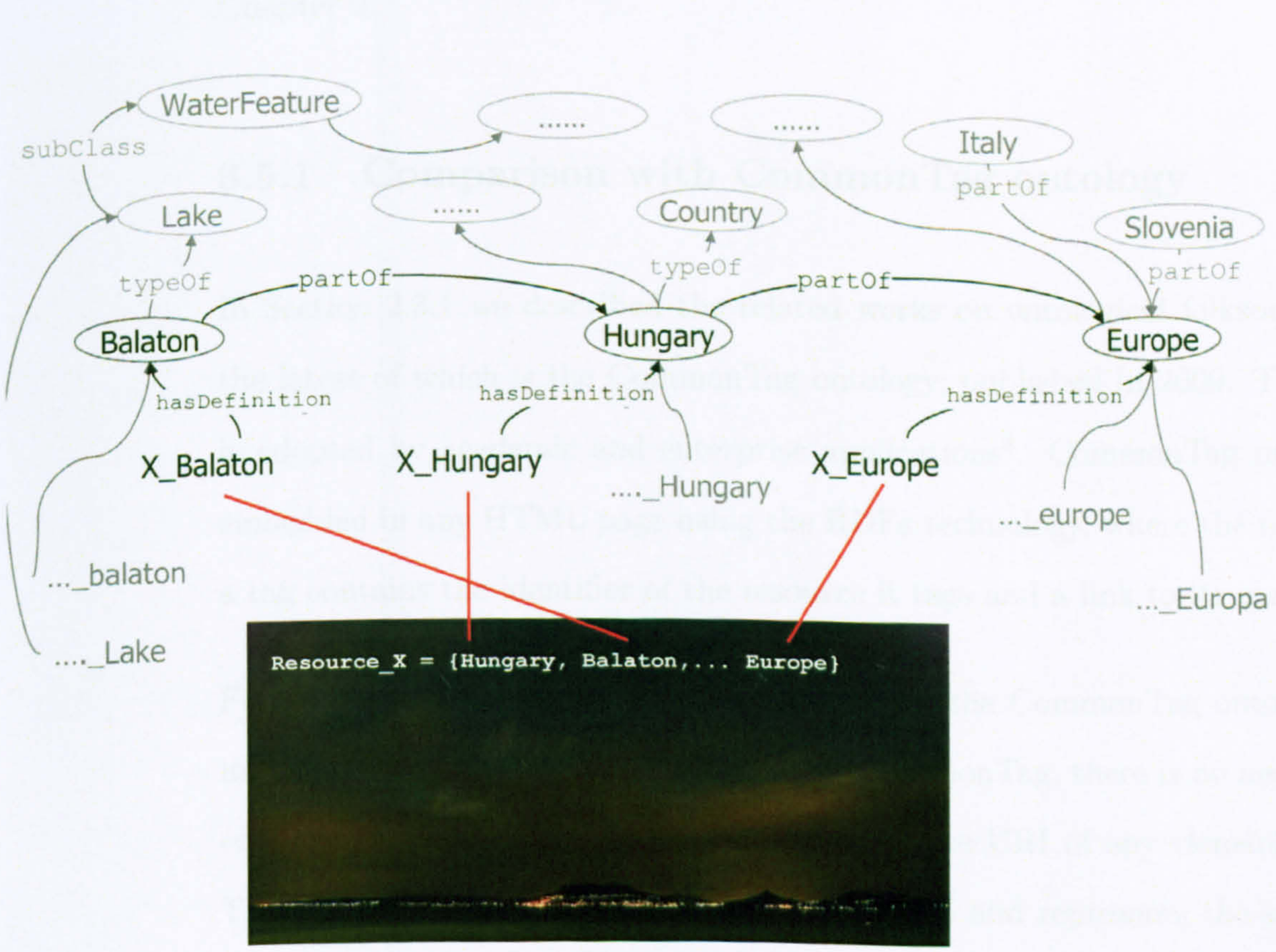


Figure 3.8: An example of FLOR enrichment for the tagset of Resource_X

They also relate with the senses of tags from other resources, for instance, Resource$_{...}$:

$$\texttt{X\_Balaton} \overset{hasDefinition}{\longrightarrow} Balaton \overset{typeOf}{\longrightarrow} Lake \overset{hasDefinition}{\longleftarrow} \texttt{...\_Lake}$$

This explicit linking of tags, resources, senses and semantic entities can largely benefit folksonomies. First of all, each tag is explicitly defined by a sense which carries additional knowledge, such as synonyms and semantic neighbourhood. As a result, complementary knowledge is injected in the resources. For example, Resource$_X$ is associated with broader concepts such as *Lake* and *Country*, which did not exist in its tagset initially. Second, such a structure can provide association of resources that represent similar concepts. For example, all the concepts that represent lake names are explicitly related with *Lake* with the *typeOf* relationship. Using appropriate semantic querying algorithms, such a structure can provide solutions to the search problems of folksonomies. We describe our approach on querying such semantic structures in Chapter 9.

3.5.1 Comparison with CommonTag ontology

In Section 2.3.1 we described the related works on ontological folksonomy modelling the latest of which is the CommonTag ontology, published in 2009. The specification is adopted by academic and enterprise applications[8]. CommonTag metadata can be embedded in any HTML page using the RDFa technology, where the representation of a tag contains the identifier of the resource it tags and a link to its meaning.

Figure 3.9 presents a graphic representation of the CommonTag ontology, as well as its alignment with the FLOR ontology. In CommonTag, there is no need for definition of a class for resource since the latter can be the URI of any element in a webpage. The relation between tag and resource is *tagged* and represents the same relation as *flor:isTaggedWith*. In addition, there is no definition of Sense. Instead, the creator of

[8] `http://www.commontag.org/Applications`

the web page is responsible for assigning to the tag an explicit URI of a concept from a Knowledge Source that represents the meaning of the tag. The relation used for this is *means* which is aligned to the *flor:hasDefinition*.

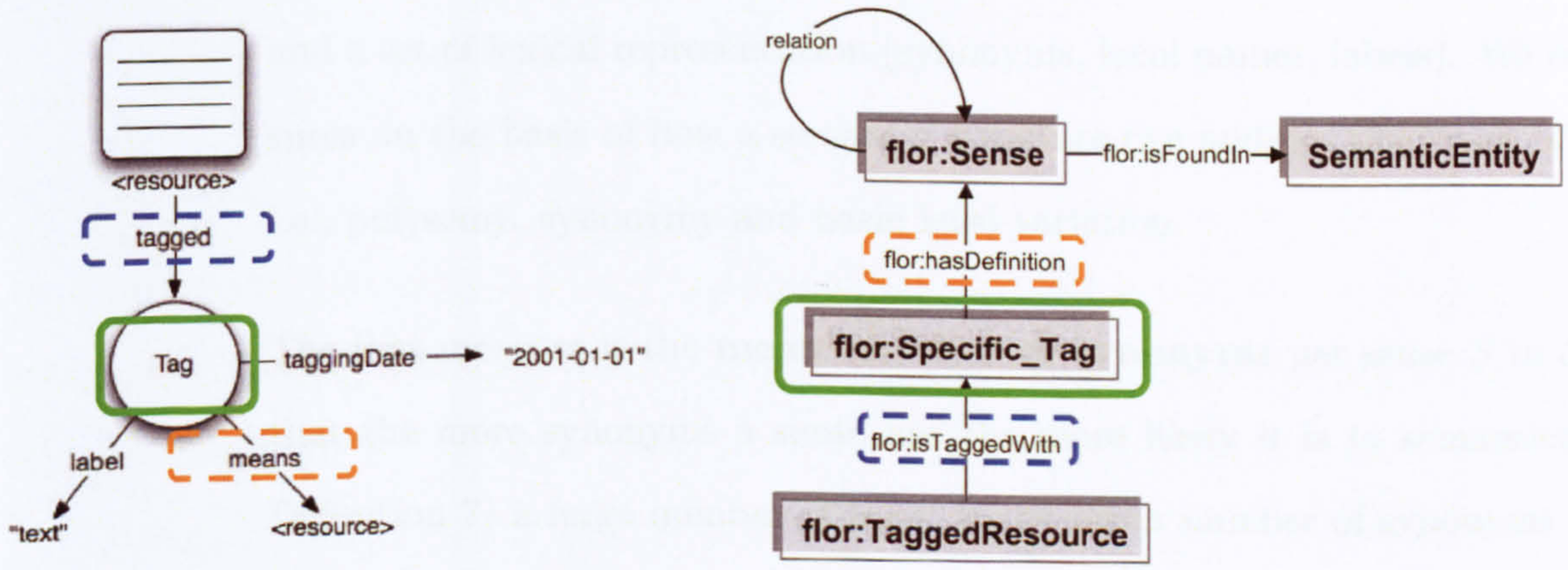


Figure 3.9: Alignment of the FLOR core ontology with the Common Tag ontology

3.6 Evaluating the Semantic Enrichment of tagspaces

The measures introduced in this section are used to asses semantic structures in terms of appropriate representation of tagspaces and improvement of search. In addition, we define measures that evaluate the performance of the enrichment algorithms in terms of tagspace coverage.

3.6.1 Evaluation Measures for Sense Structures

We define a set of measures, which we use to statistically evaluate a given **sense space**, $\mathcal{S}$, in terms of appropriate representation of a tagspace, $\mathcal{T}$. A sense space, also called **sense structure**, is a set of senses and their relations, a semantic structure, as depicted on the right hand side of Figure 3.1 and in Figure 3.8. The provenance of the sense

space (WordNet or ontologies) does not affect the calculation of the following measures. We define them based on the properties of senses (subsenses, supersenses, synonyms) however, the same measures can apply to entities from WordNet or ontologies with subordinate (hyponym, subclass) and superordinate (hypernym, superclass) entities, and a set of lexical representation (synonyms, local names, labels). We define the measures on the basis of how a semantic structure can address the issues of folksonomies, i.e., polysemy. synonymy and basic level variation.

The first measure is the **mean number of synonyms** per sense S in $\mathcal{S}$. We assume that the more synonyms a sense has the more likely it is to semantically cover (see Definition 7) a large number of tags. If the mean number of synonyms per sense in a sense structure is higher than 1, it means that the structure can potentially address the problem of tag synonymy, where different tags have the same meaning.

$$\overline{|syn(\mathcal{S})|} = \frac{\sum^{S \in \mathcal{S}} |syn(S)|}{|\mathcal{S}|} \tag{3.2}$$

For example, in Figure 3.1 $\overline{|syn(\mathcal{S})|} = \frac{1+1+2+2+1}{5} = 1.4$

With regards to tag polysemy, the more candidate senses $\mathcal{S}$ provides for one tag $\mathtt{t}$, the more likely it is to capture all its possible meanings in the tagspace. Therefore, we then define the **mean number of senses** per tag in a given tagspace $\mathcal{T}$.

$$\overline{|senses(\mathcal{T})|} = \frac{\sum_{i=0}^{|\mathcal{T}|} |senses(\mathtt{t_i})|}{|\mathcal{T}|} \tag{3.3}$$

This measure is meaningful for generic, rather than specific tags because it represents the potential for correct sense assignment. For the generic tags of the tagspace this measure represents the mean number of candidate senses per tag. For example, in Figure 3.1:

$$\overline{|senses(\mathcal{T})|} = \frac{|senses(\mathtt{fruit})+senses(\mathtt{dessert})+senses(\mathtt{apple})+senses(\mathtt{cake})+senses(\mathtt{ipod})}{5} = 1.2$$

Finally we define the **mean numbers of subsenses** and **supersenses** in a sense space $\mathcal{S}$ as:

$$\overline{|sub(\mathcal{S})|} = \frac{\sum^{S \in \mathcal{S}} |sub(S)|}{|\mathcal{S}|} \tag{3.4}$$

$$\overline{|sup(\mathcal{S})|} = \frac{\sum^{S \in \mathcal{S}} |sup(S)|}{|\mathcal{S}|} \tag{3.5}$$

In Figure 3.1 $\overline{|sup(\mathcal{S})|} = \frac{0+1+0+2+1}{5} = 0.8$. Intuitively, the higher the number of supersenses and subsenses, the higher the probability that the structure can address the issue of basic level variation. In the previously mentioned example of searching for **{lake, europe}** (Section 3.3), the more subordinate senses associated with *Europe* (e.g., areas, countries and so on), the more tags are likely to be associated with these. Therefore by having a larger number of subsenses and supersenses we may obtain a better representation of the tagspace and higher connectivity of the tags.

These measures provide quantitative assessment of a sense structure $\mathcal{S}$. We instantiate these measures in Chapters 6 and 8 where we evaluate the output of the enrichment algorithms. Note that the enrichment algorithm presented in Chapter 4 does not make use of the concept of sense (Definition 6) nor does it produce an interconnected sense space, therefore these measures are not used for the evaluation of FLOR-1.

Additionally, we define the percentage of semantically covered tags from a sense structure $\mathcal{S}$ originating from a Knowledge Source. We define the semantic overlap of the tagspace $\mathcal{T}$ and the structure $\mathcal{S}$ as $\mathcal{T}_{\text{SS}} \subseteq \mathcal{T}$:

$$\mathcal{T}_{\text{SS}} = \bigcup_{\mathtt{t} \in \mathcal{T}} \mathtt{t} : \exists\, S \in\ \mathcal{S} : Dfn(\mathtt{t}, S) = 1 \tag{3.6}$$

which is the set of tags $\mathtt{t}$ that are defined by one sense S of the sense space $\mathcal{S}$ (or are defined by one semantic entity of a Knowledge Source). Note that this refers to explicit tags as for tag $\mathtt{t}$ there is a sense S for which $Dfn(\mathtt{t}, S) = 1$. The percentage of semantically covered tags from a sense structure $\mathcal{S}$ is represented by the measure of

semantic coverage:

$$covs(\mathcal{T}, \mathcal{S}) = \frac{|\mathcal{T}_{\text{SS}}|}{|\mathcal{T}|} \tag{3.7}$$

These measures reflect the number of tags that are semantically covered from a Knowledge Source and are only dependent on the characteristics of the Knowledge Source and the characteristic of the tagspace. In particular, semantic coverage depends on the existence of an entity (or sense) that defines the meaning of a tag in a given tagset. For example in Figure 3.1, the sense structure provides 100% semantic coverage for the tagspace of the three given resources because there is one sense that defines the meaning of each tag.

We define the measure of **normalised coverage** in order to evaluate the percentage of assigned senses to tags by an enrichment algorithm, A. In other words, what percentage of tags that are semantically covered by the Knowledge Sources is correctly covered by the algorithm. If we consider the number of tags correctly covered by the algorithm to be $\mathcal{T}_{\text{A}}$ then the normalised coverage of tags by algorithm A is:

$$covn(\mathcal{T}, \mathcal{S}, A) = \frac{\mathcal{T}_{\text{A}}}{\mathcal{T}_{\text{SS}}} \tag{3.8}$$

This measure reflects the performance of the enrichment algorithms and is independent of the bias introduced by Knowledge Sources' or folksonomies' characteristics. Indeed, $\mathcal{T}_{\text{SS}}$ comprises only of the tags $\mathtt{t}$ that can be enriched as there is a sense $S \in \mathcal{S}$ for which $Dfn(\mathtt{t}, S) = 1$. In the example of Figure 3.1 $\mathcal{T}_{\text{SS}} = 6$. Consider an enrichment algorithm A, which succeeded to assign senses only to 5 out of the 6 tags of the tagspace, then the normalised coverage would be:

$$covn(\mathcal{T}, \mathcal{S}, A) = \frac{5}{6} = 0.83 \tag{3.9}$$

3.6.2 Evaluation Measures for Search

In this section we define the measures with which we evaluate the improvement of search when using semantic structures from the perspective of synonymy, polysemy, and basic level variation. We assume that for a given tagspace $\mathcal{T}$ there is a sense structure, $\mathcal{S}_{\text{KS}}$ (created using Knowledge Source KS). Consider keyword $\mathbf{k}$ used for search in the traditional keyword matching paradigm supported from the existing folksonomy systems; only results tagged with k will be returned:

$$\text{Query Results } (\mathbf{k}) = res(\mathbf{k})$$

In the example of Figure 3.1 $res(\mathbf{cake}) = \text{R}_2$ while the relevant resource R_1 is excluded. The existence of a structure that defines the relation of synonymy among tags can contribute towards solving this problem. Instead of retrieving only resources tagged with $\mathbf{k}$, the resources that are tagged with its synonyms are also retrieved. The synonym tags of $\mathbf{k}$ can be extracted from the set of synonyms of the sense $S \in \mathcal{S}_{\text{KS}}$ that defines the meaning of $\mathbf{k}$. In other words:

$$\text{Query Results } (\mathbf{k}) = \bigcup_{\mathbf{t} \in\ syn(S)} res(\mathbf{t}) : S \in senses(\mathbf{k}). \text{ Which is}$$

$$\text{Query Results } (\mathbf{cake}) = \bigcup_{\mathbf{t} \in\ syn(S_3)} res(\mathbf{t}) = \{\text{R}_1, \text{R}_2\}$$

The number of Query Results depends on the number of synonyms of the sense S that defines $\mathbf{k}$, and the number of resources each synonym $\mathbf{t}$ tags, $|res(\mathbf{t})|$. The synonyms constitute the expansion of $\mathbf{k}$ in different search scenarios. The expansion may include synonyms, subordinate and superordinate terms of $\mathbf{k}$. In particular, the expansion of $\mathbf{k}$ using a sense $S \in \mathcal{S}_{\text{KS}}$ is defined as:

$$exp(\mathbf{k}, \mathcal{S}_{\text{KS}}) = \{syn(S) \cup syn(sub(S)) \cup syn(sup(S))\}, \forall S \in senses(\mathbf{k})$$

If we denote as $res(exp(\mathbf{k}))$ the resources tagged with the expansion of $\mathbf{k}$, we can calculate the normalised increase ratio for tag $\mathbf{k}$ using a sense S from $\mathcal{S}_{\text{KS}}$:

$$ninc(\mathbf{k}, \mathcal{S}_{\text{KS}}) = \frac{|res(exp(\mathbf{k})) - res(\mathbf{k})|}{|res(exp(\mathbf{k})) \cup res(\mathbf{k})|} \tag{3.10}$$

which calculates the percentage of new resources retrieved using the expansion, compared to all the results of the system. For example, $inc(\mathbf{cake}, \mathcal{S}_{\text{KS}}) = \frac{|\{R_1,\ R_2\} - \{R_2\}|}{|\{R_1,\ R_2\} \bigcup \{R_2\}|} = \frac{1}{2} = 0.5$. The mean normalise increase for a set of keywords $\mathcal{T}$ using $\mathcal{S}_{\text{KS}}$ is calculated as:

$$\overline{|ninc(\mathcal{T}, \mathcal{S}_{\text{KS}})|} = \frac{\sum_{\mathbf{t} \in \mathcal{T}} |ninc(\mathbf{t})|}{|\mathcal{T}|} \tag{3.11}$$

In the following chapters (5, 6) we evaluate the knowledge sources and the enriched tagspaces with the measure of normalised increase and the measure of precision as this is defined in classic Information Retrieval, which reflects the percentage of correct results to all the results returned by a search system.

$$Precision = \frac{|correct(R)|}{|R|} \tag{3.12}$$

Chapter 4

First Version of Folksonomy Enrichment Algorithm

In this chapter we present the first version of the FLOR enrichment algorithm that automatically aligns tags with semantic entities from online ontologies. We apply the algorithm to a randomly selected dataset from a popular folksonomy and present our results on the correctness and the average coverage of the algorithm. We also identify a set of characteristics of folksonomies, ontologies and our approach, which we further exploit for the creation of the improved version of the algorithm.

4.1 Introduction

We present the first version of the **FoLk**sonomy **O**ntology en**R**ichment algorithm, FLOR-1. FLOR-1 is aimed at transforming flat folksonomy tagspaces into rich semantic representations using semantics from freely available ontologies. Our first attempt to create semantics tagspaces, was published in [21]. There we performed an experiment aiming to identify whether the automatic enrichment of tagspaces is feasible. In particular we reused the clusters generated by Specia and Motta in [114] and ap-

plied the relation discovery algorithm presented by Sabou et.al in [108]. Despite the straightforward approach, the algorithm automatically obtained semantic relations between the tags of the clusters. The same time we identified a set of limitations which we address in the implementation of FLOR-1.

In the approach presented in [21] we used an outdated set of ontologies[1] therefore we built the semantic entity discovery mechanism of FLOR-1 using an up-to-date search engine, the "*Watson Semantic Web Gateway*" [41], which continuously indexes Semantic Web data published on the web. In addition, in an effort to improve the tag anchoring in the semantic web (the discovery of semantic entities that contain the tag in their lexical representations), we use the rich synonym collection from WordNet to semantically expand the tag prior to searching for semantic entities that define its meaning. We point out that FLOR-1 does not deal with relation discovery among tags, it rather focuses on the correct identification of the tag concept. The relation discovery is detailed in Chapter 7.

FLOR-1 takes as input resource tagsets and for each performs three basic steps as shown in Figure 4.1. First, during **Lexical Processing** the input tagset is cleaned and tags which a-priori are highly unlikely to match semantic entities are excluded. We rely on a set of heuristics to decide which tags are likely to be less useful. Second, during **Sense Definition and Semantic Expansion** we assign a WordNet synset to each tag based on its resource context and extract all relevant synonyms and hypernyms to generate a richer representation of the tag. Finally, during **Semantic Enrichment** each tag is associated to the appropriate semantic entity.

The first step of the algorithm results in the **Lexical Representations** which is a set of lexical forms for the tag, such as plural and singular forms for nouns, or various delimited types of compound tags (sanFrancisco, san.Francisco, e.t.c). The second step identifies WordNet synset for each tag, which provide related **Synonyms** and

[1]The experiment was carried out in 2006 with data from Swoogle 2005

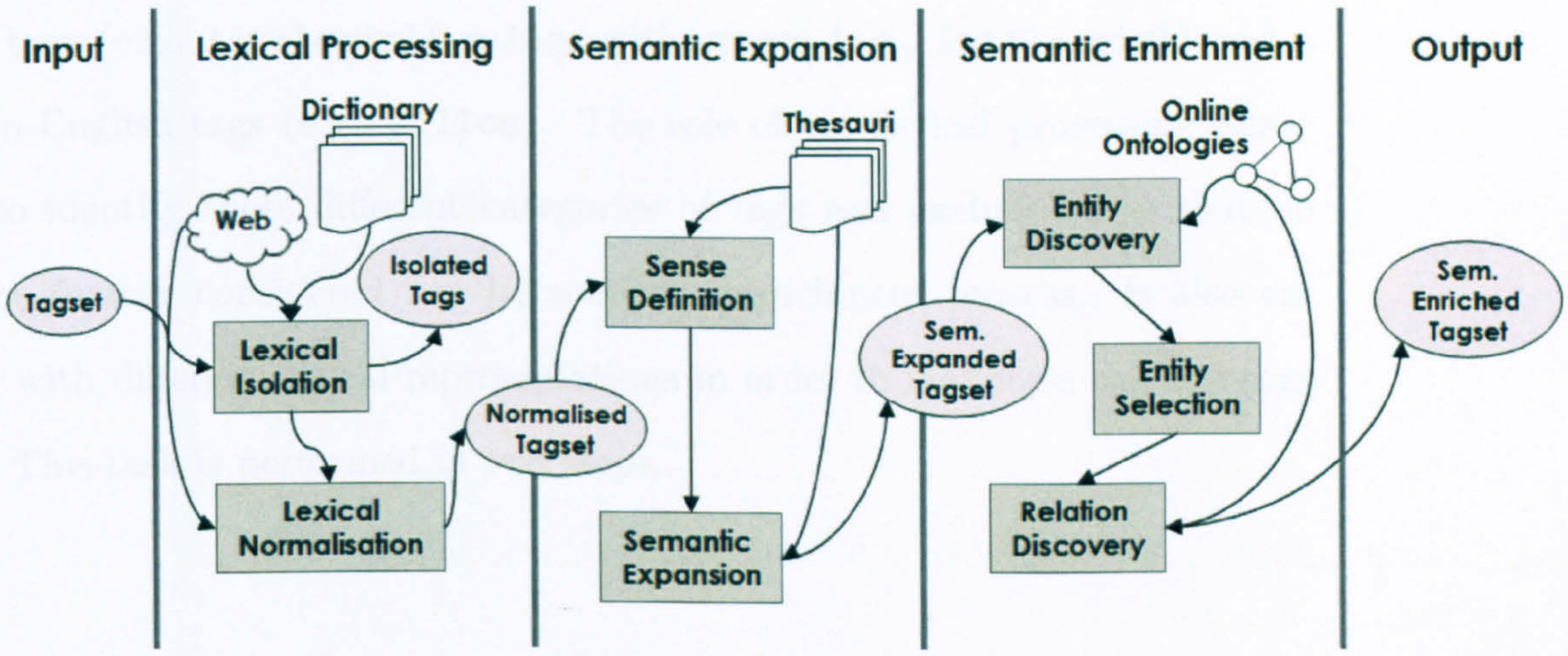


Figure 4.1: The FLOR-1 Enrichment Process

Hypernyms. The last step generates the set of the associated **semantic entities**. A tag can be associated to several relevant semantic entities because it is likely that more than one ontology may contain a valid definition for the tag.

In the following sections we describe in more detail the three phases of FLOR. In Section 4.5 we present an example of enrichment detailing each phase. Finally in Sections 4.6 and 4.7 we present our experimentation on FLOR-1 with a randomly selected dataset from Flickr the results from this study.

4.2 Lexical Processing

Due to the freedom of tagging, a wide variety of different tag types are in use. Understanding the types of tags is the first step in deciding which of them are meaningful and should be taken into account as the basis of a semantic enrichment process. Previous work [21, 51, 79] has identified different conceptual categories of tags (event, location, person), as well as tag categories that can be described by syntactic characteristics. For example, there are many tags containing special characters (e.g., `:P`), numbers (e.g., `aug07`), plurals as well as singular forms of the same word (e.g., `building`, `buildings`),

concatenated tags (e.g., `littlegirl`) or tags with spaces (e.g., `little girl`) and a number of non-English tags (e.g., `sillon`). The role of the lexical processing phase (phase 1) is to identify these different categories of tags and exclude those that do not need to be further considered for the semantic enrichment process. It also enriches the tag with different lexical representations in order to maximise the coverage in ontologies. This task is performed in two steps.

4.2.1 Lexical Isolation

The Lexical Isolation step identifies sets of tags that should be excluded as well as those that can be further processed. Currently we isolate and exclude all tags with numbers, special characters and non-English tags. The reason for excluding non-English tags is that our method exploits online ontologies, which are primarily in English. While the isolation of tags containing numbers and special characters is straightforward, the decision on the tag language is not. Although language filtering based on the existence or not of tags in an English source (for example a dictionary) can possibly rule out novel terminology, in this first version of FLOR we use WordNet for this task. WordNet is also used in the subsequent phase for the purposes of semantic expansion.

4.2.2 Lexical Normalisation

The Lexical Normalisation step aims to solve the incompatibility between different naming conventions used in folksonomies, ontologies and thesauri such as WordNet. This phase produces a set of possible **Lexical Representations** for each tag aiming to maximise its coverage by different Knowledge Sources. For example, the compound tag `santabarbara` in folksonomies appears as *Santa-Barbara* or *Santa+Barbara* in various ontologies and as *Santa Barbara* in WordNet. However, as the lexical anchoring to these resources is a quite complex problem, we try to address it by producing all

the possible lexical representations for each tag such as: {santaBarbara, santa.barbara, santa_barbara, santa barbara, santa-barbara, santa+barbara, ...}. We do so by utilising a spelling service[2] to break down the compound terms, and then use a list of delimitation patterns to produce the different formats of these terms.

4.3 Sense Definition and Semantic Expansion

Due to the phenomenon of polysemy, the same tag can have different meanings in different contexts. For example, the tag `jaguar` can describe either a car or an animal depending on the context in which it appears. As a result, in order to identify its synonyms we first need to identify its intended meaning in a certain context. In the following we describe the steps of the sense definition and semantic expansion phase (phase 2), how the tag meaning is decided using a rich sense repository, WordNet, and how this information is exploited for the alignment of the tag to ontological entities.

4.3.1 Sense Definition and Disambiguation

In this step FLOR-1 discovers the intended sense of a tag `t` in the context it appears. Since FLOR-1 deals with resource tagsets, the context of a tag in this case is the resource-based context or resource tagset T as detailed in Definition 3 of Section 3.2. We use WordNet as a sense repository and combine its hierarchy of synsets with the contextual information of `t` in order to discover a correct sense for it.

We begin by searching for WordNet synsets (senses) that define the tag and its lexical representations (generated during the lexical normalisation step). If more than one synsets are returned we exploit the contextual information of the tagset to identify the most relevant sense. We calculate the similarity between all the combinations of tags

[2] `http://search.yahooapis.com/WebSearchService/V1/spellingSuggestion`

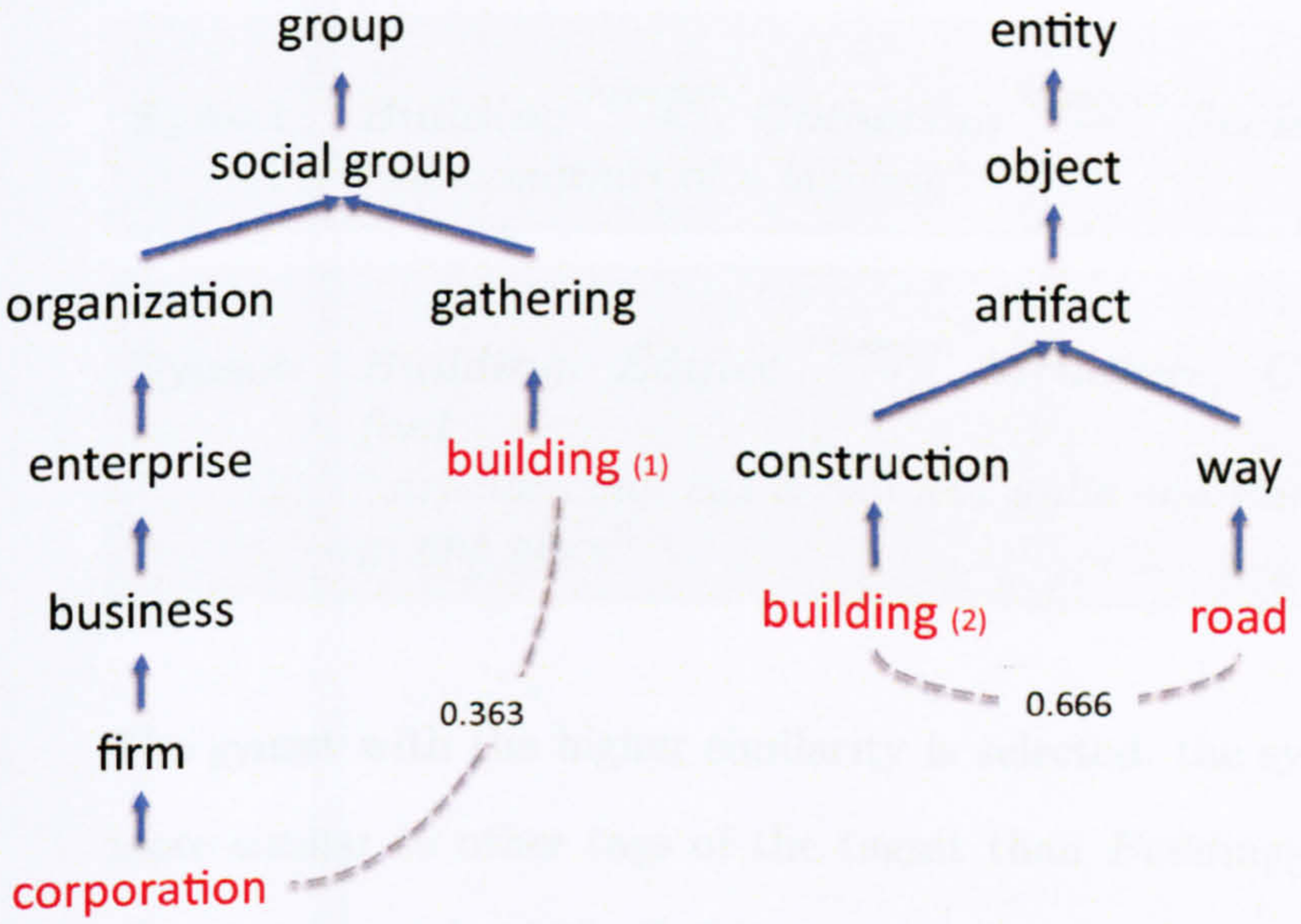


Figure 4.2: An example of the Wu and Palmer similarity measure

in the tagset using the Wu and Palmer similarity formula [125] (see Formula 4.1) on the WordNet graph. The similarity degree between two senses S_1 and S_2 is calculated based on the distance of their lowest common ancestor from the root of the hierarchy (N_3) and their distances from this ancestor (N_1 and N_2). The result for each calculation is a pair of senses and a similarity degree for these two.

$$Sim(S_1, S_2) = \frac{2 * N_3}{N_1 + N_2 + 2 * N_3} \tag{4.1}$$

For example, consider the tag `building` in the tagset {`building, corporation, road, england`}. After discovering all possible synsets for each tag we calculate the similarity of all pairs of synsets. The similarity of `building` with `england` is zero due to the lack of connection between them in the WordNet hierarchy. However, the similarities of `building` with the other two tags are calculated in the manner presented in Figure 4.2. For each pair, their nearest common ancestor is selected and the WordNet hierarchy is considered up to the most generic categories *group* and *entity*. The two synsets for building are:

Synset: ***Building*** $\overset{hyponym}{\rightarrow}$ ***Gathering*** $\overset{hyponym}{\rightarrow}$ ***Social Group***
(1) "*the occupants of a building*"

Synset: ***Building, Edifice*** $\overset{hyponym}{\rightarrow}$ ***Structure, Construction*** $\overset{hyponym}{\rightarrow}$ ***Artefact***
(2) "*structure that has a roof and walls and stands more or less permanently in one place*"

The synset with the higher similarity is selected, the synset for $Building_{(2)}$ since it is more similar to other tags of the tagset than $Building_{(1)}$. The synsets for *Road* and *Corporation* with which $Building_{(1)}$ and $Building_{(2)}$ were compared are:

Synset: ***Road*** $\overset{hyponym}{\rightarrow}$ ***Way*** $\overset{hyponym}{\rightarrow}$ ***Artefact***
"*an open way (generally public) for travel or transportation*"

Synset: ***Corporation*** $\overset{hyponym}{\rightarrow}$ ***Firm*** $\overset{hyponym}{\rightarrow}$ ***Business*** $\overset{hyponym}{\rightarrow}$ ... $\overset{hyponym}{\rightarrow}$ ***Social Group***
"*a business firm whose articles of incorporation have been approved in some state*"

If a tag has low similarities when compared to all the other tags in its cluster, then it is assigned to the most popular WordNet sense. For example, this is the case for the tag `england`, which is assigned the most popular synset[3]:

Synset: ***England*** $\overset{hyponym}{\rightarrow}$ ***European Country***
"*a division of the United Kingdom*"

4.3.2 Semantic Expansion

Once the correct WordNet synset is assigned to each tag FLOR-1 expands the tag description by including its synonyms and hypernyms. In previous example, $Building_{(2)}$

[3]In this case there is only one synset for this term

was selected as the most appropriate sense for `building` in the context of {`building, corporation, road, england`}. Then the semantic expansion associates the tag a set of synonyms: {`edifice`} and a set of hypernyms: {`structure, construction, artefact`}. These provide a richer description of the tag and are used to improve its matching to semantic entities.

4.4 Semantic Enrichment

In this phase FLOR-1 identifies the relevant semantic entities for each tag. The final output is produced by this phase and it is a set of tags enriched with the appropriate semantic entities and their semantic neighbourhood. We use the Watson API as an access point to online ontologies. We query Watson with the lexical representations (produced by the lexical normalisation) and the synonyms (produced by the semantic expansion) of a tag. All ontological Classes and Individuals returned by Watson that can potentially describe the meaning of the tag are considered. Then, we filter these entities using the hyponyms assigned to a tag during the sense disambiguation and semantic expansion phase. An example is presented in Figure 4.3 where three entities were discovered for the tag `building`.

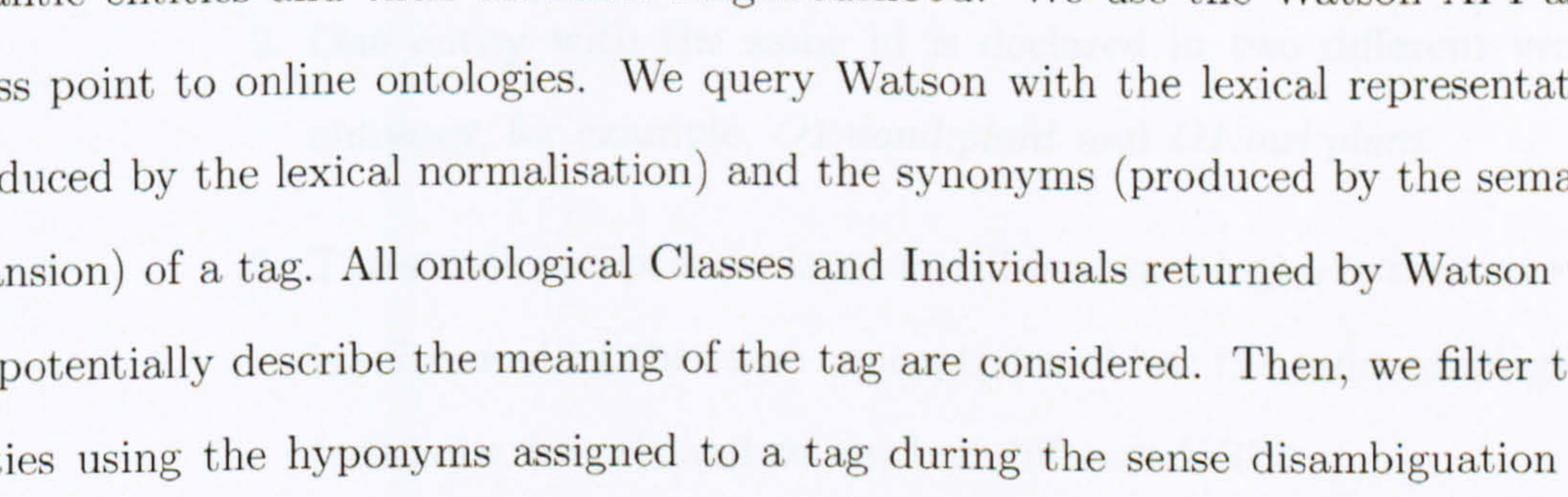


Figure 4.3: An example of selecting Semantic Entities using Hypernyms

We only select the semantic entities whose ontological parents match one (or more) of

the hypernyms. For example, because the hypernyms of `building` are {`structure`, `construction`, `artefact`} the only semantic entities selected are A and C because their superclasses flexibly match the hypernyms of `building`[4]. The qualified entities are then subjected to clustering before they are assigned to the tags.

4.4.1 Entity Clustering

By using multiple ontologies, the same concept may be defined more than once thus leading to different types of redundancies such as[5] :

1. Redundancy of the same entity. Several ontologies declare the same URI.
2. One entity with the same id is declared in two different versions of the same ontology, for example, *O1.daml:plant* and *O1.owl:plant*.
3. The same concept is declared in different ontologies in the same manner, namely it is subsumed by the same concept(s) and has the same ontological neighbourhood (relations, literals and so on) but different URI.
4. The same concept is defined in different ontologies by two different entities with different neighbourhood, for example,

Class: $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{GroceryProduce}$
(Ontology1) $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{TropicalFruit}$

Class: $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{Tree\text{-}Fruit}$
(Ontology2) *"an elongated yellowish fruit which grows on palm trees"*

[4]If the returned entity is a class, the ontological parent is its superclass. If it is an individual, the ontological parent is the class which the entity instantiates.

[5]Although the distribution of these different types of redundancies varies for different entities and a detailed experiment should be conducted to determine the exact number of such occurrences in a given snapshot of an ontological repository, our experience with the Watson repository showed that the most frequent cases of entity redundancies are 1 and 2. Almost 70% of redundancies refer to either the same URI described in different ontologies or the same entity with different base URI is described in different versions of the same ontology.

To reduce the number of redundant entities, we perform an integration process similar to the one described in [117]. The goal of this process is to set sufficiently similar semantic descriptions of entities together and merge them into a new description, a cluster of entities, which represents a single meaning. The algorithm is repeated until all obtained entities are sufficiently different from each other. To compute the similarity between two entities we compare their semantic neighbourhoods (superclasses, subclasses, disjoint and equivalent classes and named relations) as well as their lexical information (localnames, labels). The similarity $Sim(e_1, e_2)$ for two entities e_1 and e_2 is computed as:

$$Sim(e_1, e_2) = W_L \times Sim_L(e_1, e_2) + W_G \times Sim_G(e_1, e_2) \tag{4.2}$$

$Sim_L(e_1, e_2)$ is the similarity of the lexical information of the two entities computed using the Levenshtein distance metric [74]. $Sim_G(e_1, e_2)$ is the similarity of the entities' neighbourhood graphs. For example, the superclasses of e_1 are compared against the superclasses of e_2 and the same happens for subclasses and disjointness relations. This is repeated for all the neighbour entities of e_1 and e_2. The similarity among the neighbour entities is computed based on string similarity too. Because we consider the similarity of the semantic neighbourhoods more important than the similarity of the labels, we set the following restriction for the weights as: $W_l \leq W_g$.

If the similarity between two entities is higher than a threshold we merge them in one entity by integrating their neighbourhoods into one. The process is exemplified in Figure 4.4 where five semantic entities $e_{1,5}$ are compared against each other. The values in the cells $T_{i,j}$ are the similarities between the two entities, i.e., $T_{1,2}=T_{2,1} = Sim(e_1, e_2) = 0.1$.

Consider that the threshold for this example is set to 0.5. We start by performing a pair-wise comparison of the entities and observe that the pairs (e_1, e_4), (e_1, e_5), (e_2,

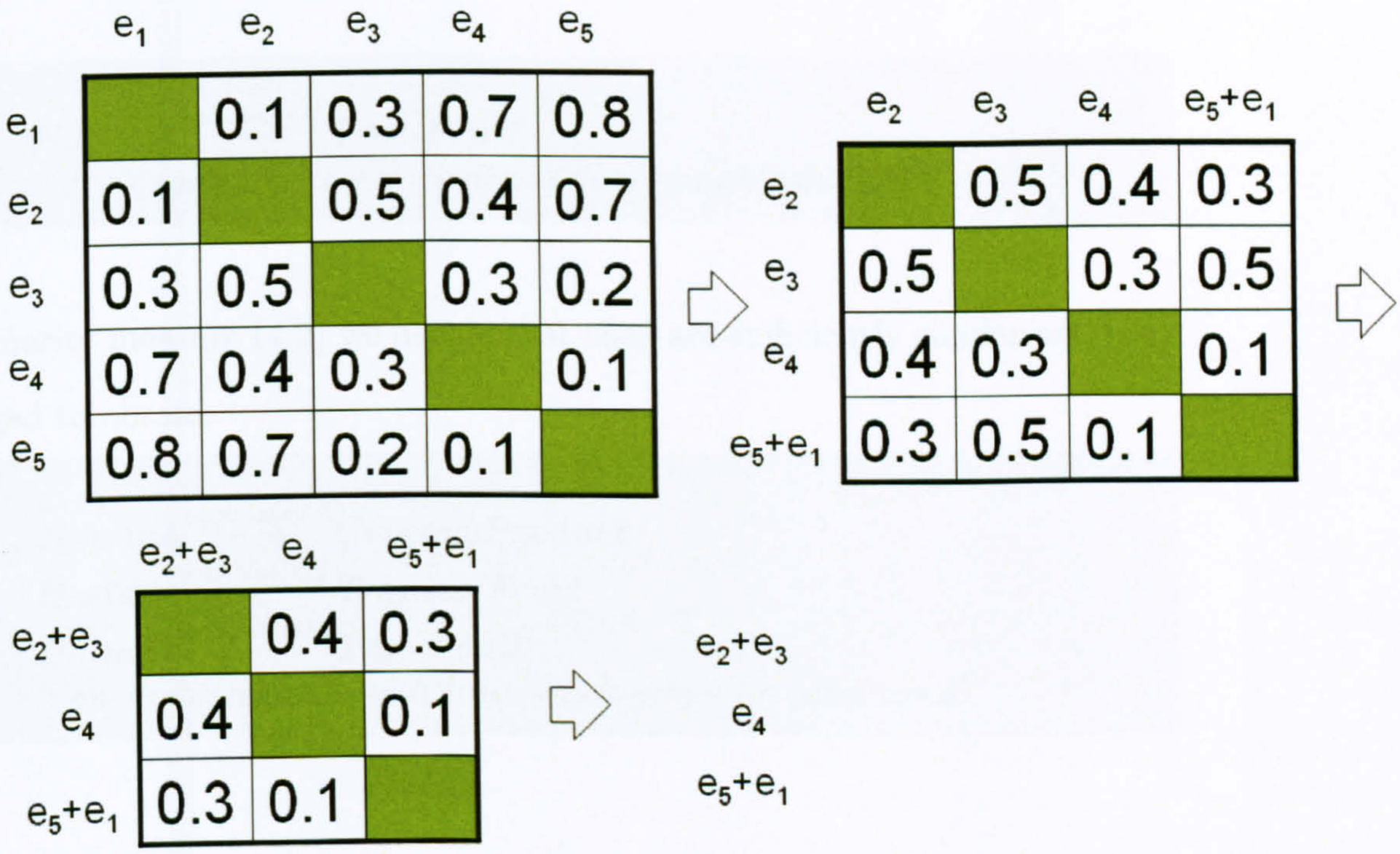


Figure 4.4: An example of entity merging strategy with threshold 0.5

e_3) and (e_2, e_5) have a similarity equal or above the set threshold. We proceed by merging the first two entities with the highest similarity, e_1 and e_5, to one entity e_1 + e_5, rearrange the table and compute the similarities between the new entity and the remaining ones[6]. This process continues until all similarities are lower than the set threshold, which implies that the obtained entities are sufficiently different. In this example, three different entities are obtained e_3+e_3, e_4 and e_1+e_5. The merging of entities is performed by collapsing their neighbourhoods. For example, Watson returns two semantic entities for banana:

Class: ***Banana*** $\overset{subClassOf}{\longrightarrow}$ ***GroceryProduce***
(Ontology1) ***Banana*** $\overset{subClassOf}{\longrightarrow}$ ***TropicalFruit***

[6]Note that in the second step, the similarity between entity e_1 + e_5 and e_2 equals to 0.3, while the similarity of e_2 with the individual entities before their merging is Sim(e_1, e_2)=0.1 and Sim(e_5, e_2)=0.7. Naturally, the merged entity e_1 + e_5 shares the features of the two component entities, therefore its similarity with the other entities (e.g., e_2) is different to the similarity of e.g., e_2 and e_1 or e_2 and e_5.

Class: $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{Tree\text{-}Fruit}$
(Ontology2) *"an elongated yellowish fruit which grows on palm trees"*

Using the similarity measure (4.2) we decide that they are sufficiently similar and they are then merged to obtain:

Class: $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{GroceryProduce}$
(Ontology1) $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{TropicalFruit}$
(Ontology2) $\boldsymbol{Banana} \overset{subClassOf}{\longrightarrow} \boldsymbol{Tree\text{-}Fruit}$
" an elongated yellowish fruit which grows on palm trees"

4.5 An Enrichment Example

In this section we present a full cycle of the FLOR-1 semantic enrichment method for the tag `lake`, which was found in the following five tagsets from Flickr:

T_1	{rush, lake, pakistan, rakaposhi, mountain,asia, kashmir, snow, sky, glacier, green, white, blue, clouds, water}
T_2	{moraine, alberta, banff, canada, lake, lac, rockies, scan}
T_3	{rising,sunlight, lake, quality, bravo}
T_4	{lake, nature, landscape, sunset, water, organisms}
T_5	{lake, finland, suomi, beach, bubbles, blue, sunlight, kids, natural}

Since the Lexical Processing phase is straightforward we do not detail it in this example (the above tagsets only contain the tags that qualified through it, i.e., initially they contained tags that were isolated by phase 1). During the second phase we queried WordNet for synsets that may define the meaning of `lake` and obtained the following three[7]:

[7]Note that synsets 2 and 3 are different synsets in WordNet despite their almost equivalent definition.

Synset: ***Lake*** $\overset{hyponym}{\rightarrow}$ ***Body of water***, ***Water*** $\overset{hyponym}{\rightarrow}$ ***Thing*** $\overset{hyponym}{\rightarrow}$ ***Entity***
(1) *"a body of (usually fresh) water surrounded by land"*

Synset: ***Lake*** $\overset{hyponym}{\rightarrow}$ ***Pigment*** $\overset{hyponym}{\rightarrow}$ ***Coloring material*** $\overset{hyponym}{\rightarrow}$ ***Material*** $\rightarrow$ ***Substance*** $\overset{hyponym}{\rightarrow}$ ***Entity***
(2) *"a purplish red pigment prepared from lac or cochineal"*

Synset: ***Lake*** $\overset{hyponym}{\rightarrow}$ ***Pigment*** $\overset{hyponym}{\rightarrow}$ ***Coloring material*** $\overset{hyponym}{\rightarrow}$ ***Material*** $\rightarrow$ ***Substance*** $\overset{hyponym}{\rightarrow}$ ***Entity***
(3) *"any of numerous bright translucent organic pigments"*

Applying the Wu and Palmer formula for the senses of `lake` and the senses of the rest of the tags in each of the tagsets we obtained variable similarities from 0 to 0.86. The zero similarities were obtained for location names such as `banff`, `pakistan`, `suomi` and for generally unrelated tags such as `quality`, `scan`, `sunlight`, `sunset`. Interestingly, `lake` returned a value of zero for the tags `glacier` and `mountain` while they should be related. The WordNet synsets for `glacier` and `mountain` are:

Synset: ***mountain, mount*** $\overset{hyponym}{\rightarrow}$ ***Natural Elevation*** $\overset{hyponym}{\rightarrow}$ ***Geological Formation*** $\overset{hyponym}{\rightarrow}$ ***Object***
"a land mass that projects well above its surroundings; higher than a hill"

Synset: ***glacier*** $\overset{hyponym}{\rightarrow}$ ***Ice mass*** $\rightarrow$ ***Geological Formation*** $\overset{hyponym}{\rightarrow}$ ***Object***
"a slowly moving mass of ice"

They are both hyponyms of *Geological formation* which is a hyponym of *object* while *Lake* is a hyponym of *Body of water* which is a hyponym of *Thing*. Although a hyponymy relation between *Lake* and *Geological formation* is expected, in the hierarchy of WordNet such relation does not exist. Furthermore *Glacier* is a hyponym of *Ice mass* but there is no subsumption relation between *Ice mass* and *Ice* or *Water* that would

allow for a connecting path between *Lake* and *Glacier*. WordNet's relations hierarchy is not sufficient for disambiguation in this case.

The highest similarity, 0.86, for `lake` was obtained with the tag `water`, because Synset 1 of *Lake* is related to *Body of water* (Synset 2 of *Water*) with a direct hyponymy relation. Note that, for most of the tagsets the first sense of *Water*, *Liquid*, is selected as this is the most common sense in which the tag `water` is used.

Synset: ***Water, H2O*** $\overset{hyponym}{\rightarrow}$ ***Binary Compound***
(1) ***Water, H2O*** $\overset{hyponym}{\rightarrow}$ ***Liquid***
"*binary compound that occurs at room temperature as a clear colourless odourless tasteless liquid*"

Synset: ***Body of water, Water*** $\overset{hyponym}{\rightarrow}$ ***Thing***
(2) "*the part of the earth's surface covered with water*"

lake

Lexical Representations	Synonyms	Hypernyms	Entities
lake		lake body_of_water water thing entity	**http://lonely.org/russia#lake** subClassOf http://lonely.org/russia#***waterway*** http://lonely.org/russia#***Lake_Baikal*** – type **http://lsdis.cs.uga.edu/proj/semdis/testbed/#lake** subClassOf http://lsdis.cs.uga..../testbed/#***Water_Feature*** subClassOf http://lsdis.cs.uga.edu/proj/semdis/testbed/#***Thing***

Figure 4.5: Enriched tag `lake` by FLOR-1

Once the correct sense is selected and the tag is semantically expanded with hypernyms (there are no synonyms for this sense of *Lake*) then the third phase of FLOR-1 queries the online ontologies through Watson and selects the semantic entities that correspond to this sense. As shown in Figure 4.5 both selected entities have the term *Lake* in their

localname and their superclass in the ontology contains one or more of the hypernyms returned by WordNet, *Water* and *Thing*, as a whole or as a compound. Note also that the selected semantic entities carry additional information about two superclasses of *Lake* (*Waterway*, *Waterfeature*) and an instance of *Lake* (*Lake Baikal*) thus further enriching the tag.

4.6 Experiments and Results

To assess the correctness of the FLOR-1 enrichment process (i.e., whether tags were linked to relevant semantic entities) we applied FLOR-1 on a Flickr data set comprised of 250 randomly selected photos with a total of 2819 specific tags. During the Lexical Isolation we removed 59% of the initial tags reducing to 1146 tags in total. We isolated 45 tags with two characters (e.g., `pb`, `ak`), 333 tags with numbers (e.g., `356days`, `tag1`), 86 tags with special characters (e.g., `:P`, `(raw → jpg)`), and 818 non-English tags (e.g., `turdus`, `arbol`). Then we filtered out the photos that exclusively contained the isolated tags (24 photos) and obtained a dataset of 226 photos with a total of 1146 tags. After running the FLOR-1 enrichment algorithm for these 226 photos, one evaluator (the author) manually checked all the assignments between tags and semantic entities.

The assignment of a semantic entity to a tag is considered correct if the concept described by the semantic entity is the same as the concept of the tag in the context of its tagset. To decide that the evaluator was given a tagset and the semantic entities linked to its tags. She evaluated each tag enrichment as "correct" if the tag was linked to the appropriate semantic entity and "incorrect" otherwise. In cases when she was not sure about the intended meaning of the tag, she rated the enrichment as "undecided". Finally, tags not associated to any semantic entity were described as "non-enriched". The results of this process are displayed in in Table 4.1.

Out of the 1146 lexically processed tags, FLOR-1 correctly enriched 281 tags and

Enrichment Result	# of Tags	Percentage
"correct"	281	24.5%
"incorrect"	20	1.7%
"undecided"	4	0.3%
"non-enriched"	841	73.4%
Total	1146	100%

Table 4.1: Evaluation of semantic enrichment for individual tags.

incorrectly enriched 20 tags thus leading to precision results of 93%. An example of incorrect enrichment is that of `square` in the context {`street`, `square`, `film`, `color`, `documentary`}. While its intended meaning is *Geographical area*, because during the disambiguation phase `square` did not return high similarity with any of the rest of the tags, the WordNet sense assigned to it was the most popular one, *Geometrical shape*.

Synset: ***Square, Foursquare*** $\overset{hyponym}{\rightarrow}$ ***Rectangle*** $\overset{hyponym}{\rightarrow}$ **....** $\overset{hyponym}{\rightarrow}$ ***Shape***
"(geometry) a plane rectangle with four equal sides and four right angles; a four-sided regular polygon"

This led to the assignment of non-relevant semantic entities namely:

Class: ***Square*** $\overset{subClassOf}{\longrightarrow}$ ***Rectangle***
(Ontology 1) *"Any Rectangle whose sides are all equal"*

Class: ***Square*** $\overset{subClassOf}{\longrightarrow}$ ***Rectangle-2D***
(Ontology 2) *"[...] Each instance of Square is a rectangle with all four sides of equal length.[...]"*

Despite this error, the rest of the tags in this tagset were correctly enriched.

FLOR-1 failed to enrich 841 tags, i.e., 73.4% of the tags (see Table 4.1). Because this is a significant amount of tags, we wished to understand whether the enrichment failed because of FLOR's low coverage enrichment or because most of the tags have no equivalent coverage in online ontologies. To that end we selected a random 10%

of the 841 tags (85 tags) and manually identified appropriate semantic entities using Watson and taking into account the context(s) of the tags in the tagset(s) they appear. Out of the 85 tags we manually enriched 29. We therefore estimate that the number of tags that could have been enriched by FLOR-1 (i.e., those for which an appropriate semantic entity exists) is approximately 287. Thus, taking into account that the overall number of tags that should be correctly enriched was 568 (281+287)[8] but only 281 were enriched by FLOR-1[9] this leads to an approximate normalised coverage for FLOR-1, M3.8:

$$covn(\mathcal{T}, \mathcal{S}, \text{FLOR-1}) = 49\%$$

where $\mathcal{T}$ represents the tagspace of the experimental dataset and $\mathcal{S}$ the online ontologies in Watson as a Knowledge Source. While this is quite a low enrichment percentage, these results are highly superior to the ones we have obtained in previous experiments [21] where we did not perform semantic expansion and we directly searched for semantic entities for the tags without relying on WordNet as an intermediary step. Indeed, the WordNet sense definition and expansion of the tags with synonyms and hypernyms (FLOR-1 phase 2) increased the tag discovery in the Semantic Web thus having a positive effect on the coverage of tags to ontologies.

FLOR-1 failed to enrich the above 29 tags due to the following reasons. The majority of the failures (55%) was due to different definition in terms of superclasses in WordNet and in online ontologies For example, the definition of `love` in WordNet and the relevant entity found in the Semantic Web are:

Synset: ***Love*** $\overset{hyponym}{\rightarrow}$ ***Emotion*** $\overset{hyponym}{\rightarrow}$ ***Feeling*** $\overset{hyponym}{\rightarrow}$ ***Psychological feature***
"*a strong positive emotion of regard and affection*"

[8]This is equivalent to $\mathcal{T}_{\text{SS}} = 568$, see M3.6 in Section 3.6.1
[9]This is equivalent to $\mathcal{T}_{\text{A}} = 281$

Class: ***Love*** $\overset{subClassOf}{\longrightarrow}$ ***Affection***
"Love is a collection [..]. Specialized forms of Love are Love-Romantic, platonic love, maternal love, infatuation, agape, etc."

Although both these definitions refer to the same sense, and additionally the superclass *Affection* belongs to the gloss of *Love* in WordNet, they were not matched because *Affection* does not appear as a hypernym of *Love*.

A further 24% of the tags not connected to any semantic entity were assigned to the wrong WordNet synset during phase 2. For example, `bulb` referring to `light bulb` in its tagset is assigned the incorrect synset:

Synset: ***Bulb*** $\overset{hyponym}{\rightarrow}$ ***Stalk, Stem*** $\overset{hyponym}{\rightarrow}$ ***Plant organ***
"modified bud consisting of a thickened globular underground stem serving as a reproductive structure"

The rest of the tags are unenriched due to failures in anchoring them into appropriate semantic entities. This is because, despite the lexical enrichment phase FLOR uses strict string matching.

For 4 tags the evaluator could not determine whether the enrichment was correct or incorrect (Table 4.1). This is because the meaning of the tag was unclear even when considering its context and the actual image it is assigned to. For example, in the photo of Figure 4.6 the meaning of the tag `volume` is unclear. In the second phase of FLOR-1 the tag was expanded with the hypernyms *Measure* and *Abstraction* from the most popular synset of WordNet for `volume`:

Synset: ***Volume*** $\overset{hyponym}{\rightarrow}$ ***Measure*** $\overset{hyponym}{\rightarrow}$ ***Abstraction***
"the amount of 3-dimensional space occupied by an object"

Then, it was related to the semantic entity:

Figure 4.6: A case of "undecided" Enrichment

Class: $\textbf{\textit{Volume}} \overset{subClassOf}{\longrightarrow} \textbf{\textit{Measure}}$
"The measure of the physical space of any 3-D geometric object"

As the meaning of the tag was not clear for the evaluator, she evaluated it as "undecided". More generally, there are several cases when tags only make sense to their author (and maybe to his social group) and thus are difficult to enrich.

4.7 Lessons Learnt

We presented the FLOR-1 enrichment algorithm and the experiments we conducted on a subset of Flickr photos. We enriched approximately 49% of the tags (normalised coverage) with a precision of 93%. Compared to our previous efforts to map tags

to Semantic Web Entities, without previously expanding them with synonyms and hypernyms (23% in [21]), this is a significant improvement. Analysing the results we identified a number of issues to be resolved in order to enhance the performance of FLOR-1.

As indicated by the results in Section 4.6, the cases of incorrect enrichment and lack of enrichment were mainly caused by failure in the **Sense Definition and Semantic Expansion** phase. In particular the following issues need to be addressed in order to correct the errors and enhance the performance of this phase. First, it is essential to extend the tag similarity measure to also exploit other relations rather than only subsumption. This flaw was exemplified in the case of `lake` and `glacier` which were considered unrelated based on the hierarchical structure of WordNet (Section 4.5). Second, in the example of `square` co-occurring with `street`, the incorrect sense definition for `square` caused further incorrect enrichment (Section 4.6). One of the possible solutions to this is to perform context expansion exploiting tag co-occurrence. For example, expanding the {`square`, `street`} tagset with their frequently co-occurring tags, for example {`building`, `park`}, can increase semantic relatedness between the tags and potentially lead to correct mappings from tags to correct senses.

To conclude, we present the issues highlighted from this experimentation with FLOR-1. To increase the coverage and correctness of assigning tags to semantic entities it is required to:

- **L4.1:** Identify alternative relatedness measures among the tags of a tagset.
- **L4.2:** Utilise tag co-occurrence where semantic measures fail.
- **L4.3:** Reconsider the use of WordNet as a source for semantic expansion.

In Chapter 5 we present an evaluation of this enrichment algorithm from a search perspective with the help of a user study.

Chapter 5

Searching Enriched Tagspaces: Initial Experiments

In this chapter we present an initial experiment on querying a tagspace enriched with the first version of FLOR. We describe an algorithm for search in the enriched tagspaces and its implementation as a web application. We perform a user experiment and report on the user experience, as well as, the performance of the enriched tagspaces in search.

5.1 Introduction

In this chapter we investigate how the enriched tagspaces obtained with FLOR-1 compare against the flat tagspaces in a search scenario. In Chapter 2 we described the relevant work on folksonomy improvement and highlighted the lack of formally established evaluation benchmarks. As a result, some of the folksonomy improvement approaches [17, 85, 102] have conducted a tailored evaluation which best suits the characteristics of the problem they address and the methods they propose. For the same reasons, we perform an evaluation of FLOR-1's impact on search using a larger dataset than the one used in the evaluation described in Chapter 4. Here, we enrich a dataset

from Flickr, build a query expansion mechanism, which facilitates semantically-enabled search, and use it to perform a user experiment.

In Section 5.2 we briefly describe the enrichment of the tagspace, introduce the query mechanism and present the user interface. In Section 5.3 we describe the experiment conducted and the results we obtained. We conclude with Section 5.4 where we describe the lessons learnt from this experiment.

5.2 Method

5.2.1 Enrichment

For this experiment we used the algorithm described in Chapter 4, FLOR-1, to enrich the input tagspace. FLOR-1 takes as input a set of tagsets corresponding to resources and for each tagset T performs the following steps. At first it eliminates the less useful tags and lexically processes the rest. Then, each tag `t` is disambiguated by being matched to an appropriate WordNet synset according to the context of T. Using the synset's synonyms and hypernyms we identify semantic entities from online ontologies and connect them to `t`.

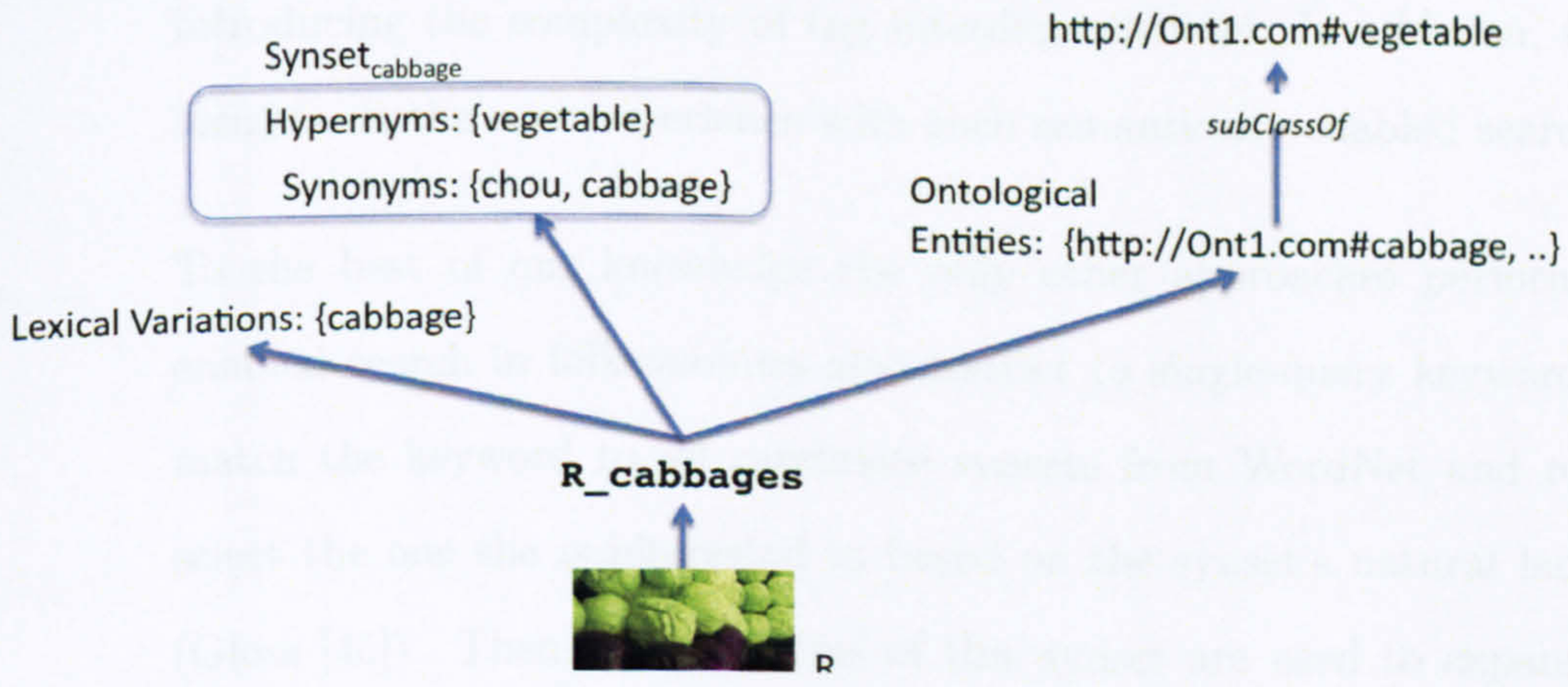


Figure 5.1: Enriched tag `cabbage` by FLOR-1

Figure 5.1 depicts the representation of the output of FLOR-1. Resource R, is tagged with specific tag **R_cabbages** which is associated with three types of entities according to the three phases of FLOR-1. First the tag is lexically enriched with a lexical variation {**cabbage**} and then it is associated with synset $\text{Synset}_{\text{cabbage}}$. From that it is semantically expanded with synonyms {**chu, cabbage**} and hypernyms {**vegetable**}. Finally it is associated with ontological entities, for example *http://Ont1.com#cabbage*. Each ontological entity is associated with other semantic entities such as:

$$http://Ont1.com\#cabbage \overset{subClassOf}{\longrightarrow} http://Ont1.com\#vegetable \tag{5.1}$$

In the next section we describe different strategies to perform search on enriched tagspaces in order to investigate the value of the enrichment algorithm.

5.2.2 Querying Strategies

We describe three search strategies, each of which exploits different characteristics of the enriched tagspace in order to obtain relevant results to a query keyword **k**. In this preliminary experiment we allow only for single-keyword queries. We pose this limitation because we want to study individual cases of tag enrichment without introducing the complexity of tag interdependencies. In addition, we want to obtain insights on the user experience with such semantically-enabled search.

To the best of our knowledge the only other approaches performing semantically-enabled search in folksonomies also restrict to single-query keywords. Lee et. al [73] match the keyword to all candidate synsets from WordNet and require the user to select the one she is interested in based on the synset's natural language description (Gloss [45]). Then the synonyms of this synset are used to expand the query. This work has not been evaluated from a user perspective. A work published at a later time than the one described in this chapter is presented by Pan et. al [96]. They intro-

duce an approach for reducing tag ambiguity in domain specific search in folksonomies (e.g., music videos in Youtube). Their approach depends on appropriately bootstrapping their expansion framework with a relevant domain ontology. They predefine a set of queries and evaluate the performance of their method by comparing the relevance of the results to each query keyword and using a measure of precision (equivalent to Measure 3.12). The relevance judgement is provided by one evaluator.

In contrast to Pan et. al, who deal with the problem of tag ambiguity (i.e., tag polysemy), we investigate how the structure of enriched tagspaces can help address the problems of synonymy and basic level variation (see Chapter 3). As a result we introduce the following search strategies.

(A) Querying with tags

This is the type of search provided by folksonomies where the set of results consists of all the resources that are explicitly tagged with query keyword **k**. In the example of Figure 5.1, **k** is compared only against the tag itself, i.e., **R_cabbages**. R is only retrieved if **k=cabbages**. We use this strategy as the baseline for comparison against the following two.

(B) Querying with synonyms and lexical variations

With this strategy we aim to increase the number of resources by dealing with the issue of synonymy. As a result, the results of this strategy are resources tagged with synonyms or variant lexical representations of the query keyword **k**. The lexical representations are obtained by the phase of Lexical Processing (Section 4.2.2) and the synonyms from the phase of Semantic Expansion (Section 4.3.2). The results of this strategy are the resources tagged with tags **t**, whose synonyms or lexical variations contain the query keyword **k**. In the example of Figure 5.1, R will be retrieved for **k={cabbage, cabbages, chou}** although it is only tagged with **cabbages**.

(C) Querying with subclasses and hyponyms

With this strategy we investigate the problem of basic level variation by returning

resources annotated with subordinate terms of the query keyword. Because not all semantically expanded tags are successfully associated with semantic entities[1] we also use the hierarchical relations of WordNet. In this strategy, the query keyword **k** is mapped against the hypernyms of the tags' synsets and the superclassess of the tags ontological entities. In this strategy the results are resources tagged with tags **t** that:

- contain the query keyword **k** in their set of hypernyms
- are associated with semantic entities which are subclasses of the entities with which **k** is matched.

In the example of Figure 5.1, R will be retrieved for **k**={**vegetable**}. In the following section we describe the user interface that supports search using these three strategies.

5.2.3 User Interface

We build a web interface implementing the above strategies, using JSP[2] on Apache Tomcat[3]. The enriched tagspaces reside on a Sesame[4] RDF repository and strategies (A) to (C) are implemented in terms of SerQL query calls to the repository.

The introductory page of the web application consists of a search box and instructions to the users. The first page of the results is displayed in Figure 5.2. The first column (A) presents the results that are explicitly tagged with the query keyword; for example, **vegetable** and represents the tag-based search in folksonomies. Column (B) presents the results that are tagged with synonyms or different lexical representations of the query term; for example, {**veggies, vegetables**} and implements strategy (B). Finally, column (C) presents the results tagged with subordinate senses of the query keyword; for example, {**legume, artichoke**} and represents querying strategy (C).

[1] see the example of **love** in Section 4.6
[2] **http://java.sun.com/products/jsp/**
[3] **http://tomcat.apache.org**
[4] **http://www.openrdf.org/**

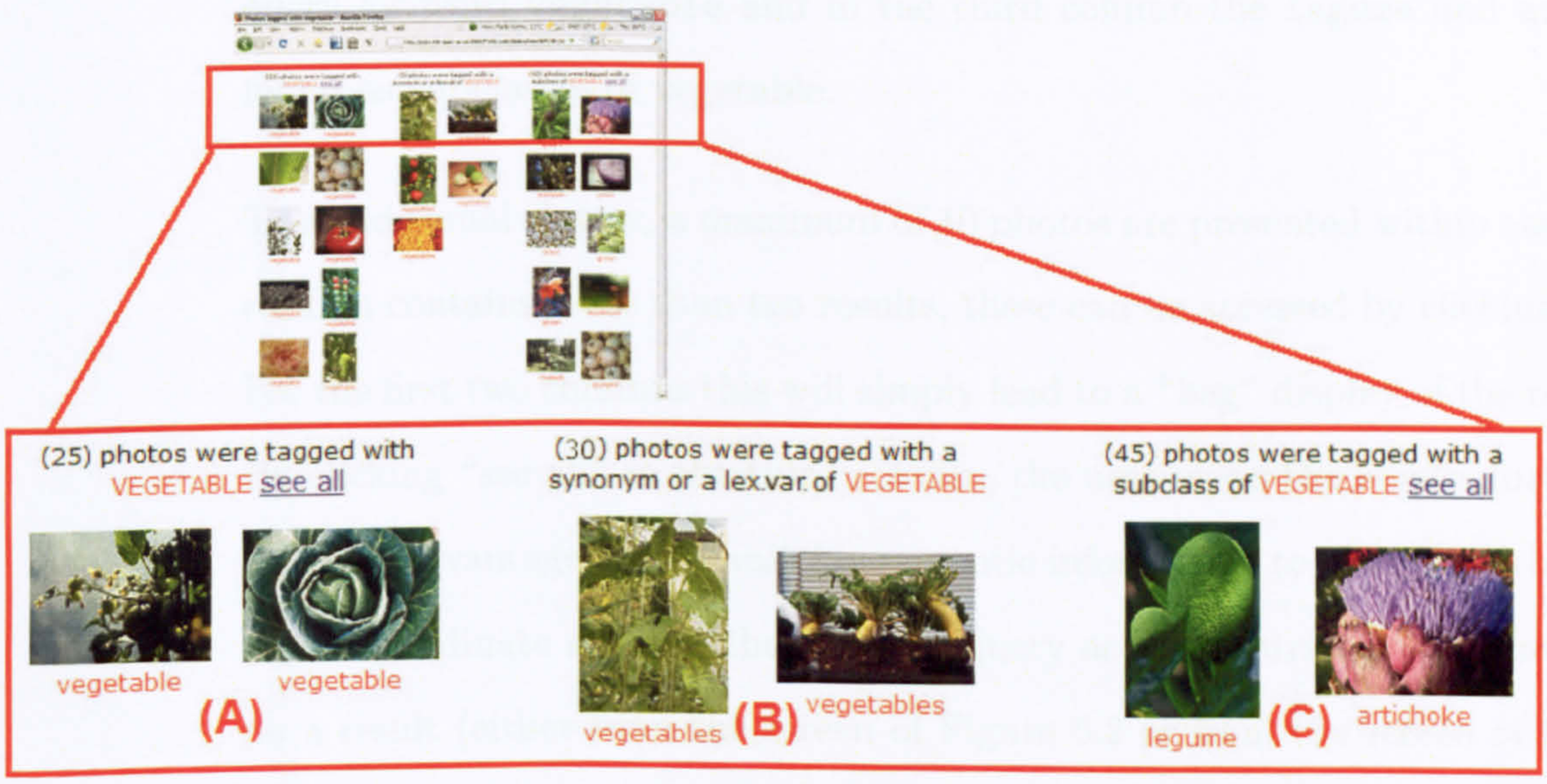


Figure 5.2: First Page - Results for `vegetable`

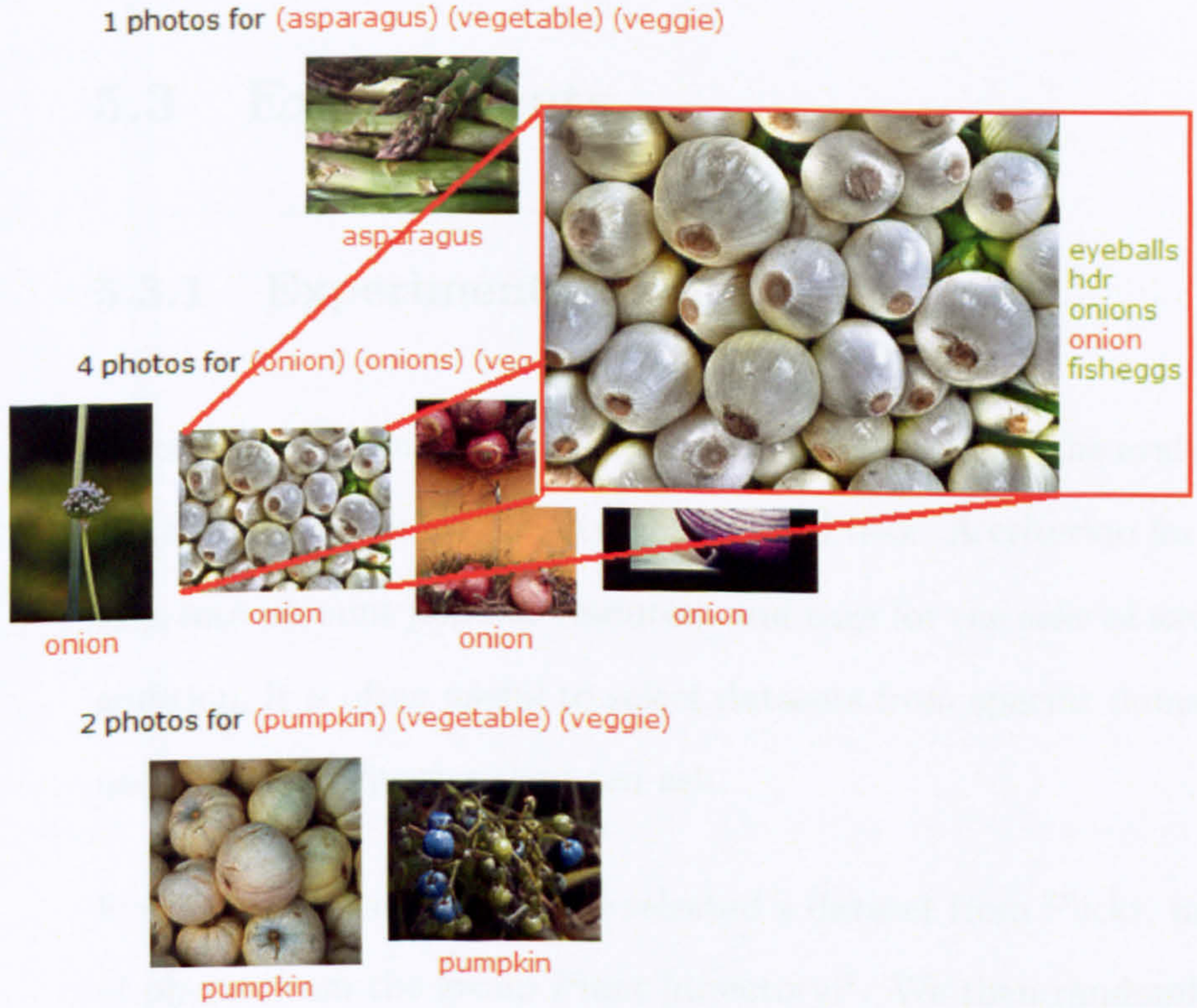


Figure 5.3: Second Page - Results tagged with subordinate concepts of `vegetable`

Under each result we present the tag which caused a resource to be returned for this query. For example, in the second column the tag `vegetables` was matched to the

query keyword **vegetable** and in the third column the **legume** and **artichoke** were found as subclasses of vegetable.

To avoid visual clutter, a maximum of 10 photos are presented within each column. If a column contains more than ten results, these can be accessed by clicking on **"see all"**. For the first two columns this will simply lead to a "bag" display of the relevant photos. By clicking "see all" in the third column, the user views the results further organised. We take advantage of the available semantic information to categorise the results under each subordinate sense of the keyword query as demonstrated in Figure 5.3. Clicking on a result (either from the screen of Figure 5.2 or from the screen of Figure 5.3) the user obtains a larger view of the photo with all its associated tags and the tag matching the query keyword highlighted (e.g., **onion**).

5.3 Experiments

5.3.1 Experimental Setup

As previously mentioned, the selection of a dataset for the evaluation of folksonomy enrichment methods is not a straightforward task. A criterion for a good selection should take into account popular resources and tags for the sake of avoiding idiosyncrasies. In addition, it is often useful to select datasets from specific domain in order to guide the users with the queries they can ask.

For the current evaluation we selected a dataset from Flickr, more specifically a subset of photos from the group Plant [directory][5]. We then randomly selected 12233 photos with a total of 13645 generic tags (89446 specific tags). Applying FLOR-1 on this dataset we acquired the following results: 3765 tags, i.e., 27,6% of the generic tags were disambiguated and semantically expanded with (WordNet) synonyms in Phase

[5] **http://www.flickr.com/groups/plantdirectory/** 5.943 members and 63.454 resources on 24-07-2008

2. Out of these 670 tags, i.e., 4% of the total tags were linked to Semantic Web Entities resulting into 9697 (79,2%) enriched photos. The enrichment precision of the algorithm (i.e., how many enriched tags were correctly enriched) was tested in the earlier evaluation of FLOR-1 in Section 4.6. In that dataset (250 photos and 2819 individual tags) 25% of the tags were enriched with 93% precision. The drop in the tag coverage (25% to 4%) can be explained by the fact that the majority of the photos belonging to the Plant [directory] group are tagged with group idiosyncratic tags such as {`ilovenature, naturesfinest, lovely1`}. The existence of idiosyncratic tags drops the coverage percentage but does not affect the enrichment of the rest of the tags in the group since they cannot contribute to the disambiguation phase by not being matched against WordNet synsets.

The percentage of enriched tags (4% during Phase 3) with semantic entities compared to the percentage of semantically expanded tags (27,6% during Phase 2) is quite low. As the goal of this experiment was to evaluate the usefulness of the enriched tagspaces to the user, it makes sense to concentrate on the structure provided by WordNet leaving out the ontological entities returned for the 4% of the tags.

5.3.2 User Study

We asked 11 users (postgraduate and postdoctoral researchers) to post at least 3 single keyword queries related to plants using the web application described in Section 5.2.3. They had to evaluate the results returned in each column and answer the questions of Table 5.1. The results for **tag-based** search (T) refer to strategy (A). (S) represents systems (B) and (C) which are obtained using the **semantically-enabled** search.

After the completion of the experiment we acquired 45 individual user queries. The following four {`aquatics, bryophytes, conkers, photosynthesis`} did not return any results neither in (T) nor in (S), as no photos from the dataset were tagged with

Q1:	What are you looking for?		
Q2:	What keyword did you use?		
Q3:	Did you find what you were looking for?	(T): 88%	(S): 88%
Q4:	Did the presentation of the results help you find what you were looking for?	(T): 77%	(S): 88%
Q5:	Are there any photos that should not be in your results?	(T): 12%	(S): 21%
Q6:	Did you find any photos with (T) that you were not able to find with (S)? And vice versa?	(T): 0%	(S): 66%

Table 5.1: User Experiment 1: Questions and Responses

them[6].

Additionally, there were nine keywords for which (S) did not return any additional results to (T). This was because FLOR-1 did not enrich the tags corresponding to these keywords. One of them is the misspelled **`funghi`** (the correct spelling is **`fungi`**). Yet, there were two photos tagged with the misspelled tag but no WordNet synsets or semantic entities exist (at the time of the experiment) for **`funghi`**. In other cases (S) did not return additional results as the query keywords did not have any synonyms (returned by (B)) or hyponyms (returned by (C))[7]. Finally for the keyword **`fish`** although it was expanded by (S) no photos were found to be tagged with its synonyms or subclasses as the domain of our dataset was restricted to plants. Yet this yielded a useful observation. The correct assignment of a semantic entity to a tag is not enough if this entity's neighbourhood does not cover other tags in the tagspace. In other words, semantic entities whose neighbourhoods are assigned to a larger number of tags in the tagspace are more useful to the enrichment process.

In the following sections we present quantitative results on normalised increase (Measure 3.11) and precision (Measure 3.12) calculated from the 32 remaining user keywords and describe the user incentives on this experiment.

---

[6]The non existence of photos tagged with them (or their synonyms) did not trigger the enrichment process thus strategies (B) and (C) could not find relevant tags `t` for these keywords.

[7]These keywords are: **`lotus`**, **`aloe`**, **`trunk`**, **`oak`**, **`stigma`**, **`boletus`**, **`wither`**

Normalised Increase and Precision

In Chapter 3 we defined the measure of normalised increase M3.11 as the percentage of additional results obtained when querying with the expansion of the query keyword over the total results (obtained with a keyword and its expansions). We calculated the additional correct results with the help of one evaluator who judged the results of the queries based on the information needs of the users (see Table 5.1: Q1) and their responses (see Table 5.1: Q3-Q6).

The light (green) bar in Figure 5.4 represents the normalised increase values of (S) for the 32 user keywords, $\mathcal{K}$, ranging from 0 to 98% and obtaining average normalised increase M3.11:

$$\overline{|ninc(\mathcal{K}, (\mathrm{S}))|} = 36\%$$

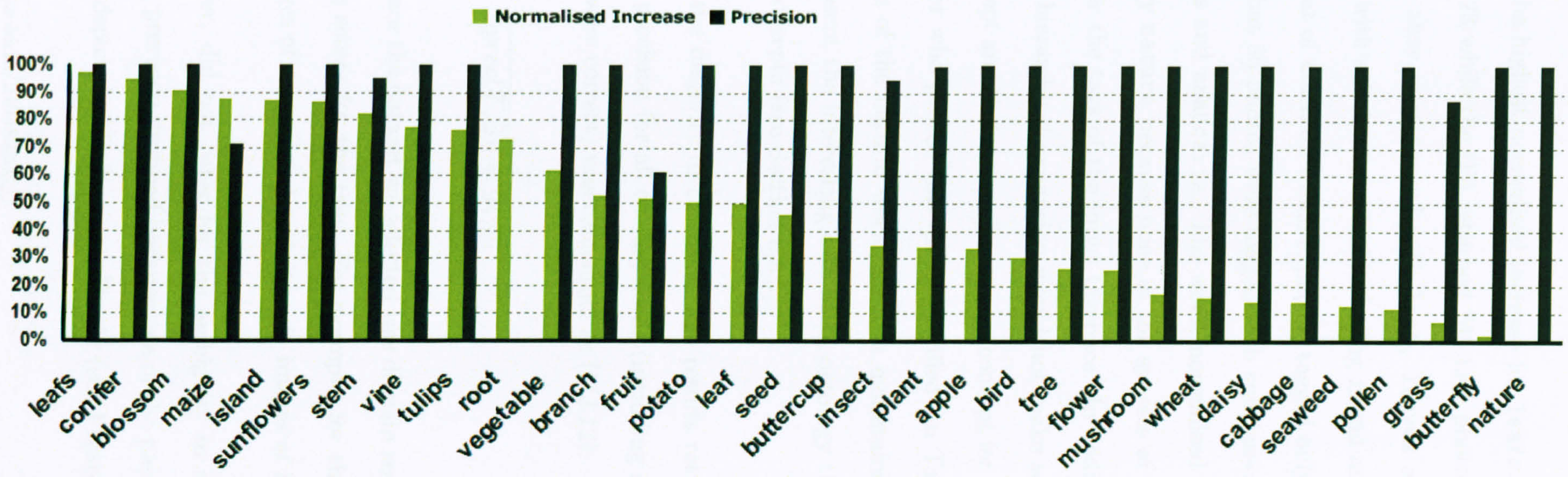


Figure 5.4: Normalised Increase and Precision of Strategies (B) and (C) on $\mathcal{K}$

We observe that (S) achieved the highest normalised increase for `leafs`. In this case the number of (T) results was 23 while the (B) returned 745 new photos that would otherwise be excluded because they were tagged with `leaf`, `leaves` and `foliage`. (C) returned 49 photos tagged with subclasses of leaf including `frond` and `rossette`. Another interesting case was that of `conifer` which explicitly tagged only two photos. (B) did not return any results but 36 photos were tagged with subclasses of `conifer` such as `pine`, `cedar`, `cypress` and `araucaria`. The worst normalised increase rate (0%) was achieved for the query `nature` because none of the synsets of WordNet for nature has hyponyms that cover the tags of the input tagspace. In addition, `nature` is one of the queries for which hierarchical expansion would not make sense. This is because *nature* is a generic concept and the notion of narrower concept for nature varies among different users. The user who posted this query specified (in Table 5.1: Q1) that she was looking for images of the natural world, such as mountains, landscapes and so on. In Chapter 6 we present the interesting case of an ontology that describes *nature* as a superclass of such concepts (see Section 6.3.3).

In Figure 5.4 is shown that in the majority of the cases, the results returned by (S) are 100% correct. The average precision for all the keywords (including `maize`, `root`, `fruit` for which not all results were correct) was calculated as (M3.12):

$$\overline{|prec(\mathcal{T}, \mathrm{S})|} = 94\%$$

This is not a surprising result since the dataset we selected was domain restricted, thus less likely to contain ambiguous senses for the tags. For example, for the ambiguous word `lotus`, there were no images of cars tagged with it, only images of flowers.

The domain restriction, however, did not cater for the ambiguity in the following case. For the query `insect`, the precision dropped because one of the photos retrieved with the hyponyms of `insect` depicted a Volkswagen Beetle (see Figure 5.5[8]). The

[8]http://www.flickr.com/photos/kazukichi/2404898062/

insect sense of *beetle* was incorrectly assigned to this image because in FLOR-1 if the WordNet-based disambiguation fails the most popular sense is assigned to the tag. This is a known limitation of the algorithm which can be eliminated by employing other Knowledge Sources, where more tags are covered and their relations e.g., between *beetle* and *car* are declared (we have implemented this improvement in the second version of FLOR Chapter 7). Yet an interesting observation emerges from this example.

The image depicts different objects which are not semantically related and its tagset ({`om1`, `tree`, `yokohama`, `plants`, `japan`, `flower`, `kanagawa`, `multipleexposure`, `car`, `beetle`}) contains different contexts. For example, {`tree`, `plants`, `flower`} define one context, {`yokohama`, `japan`, `kanagawa`} define another, and the same holds for {`beetle`, `car`}[9]. The existence of various sub-contexts (within the context of a tagset) which are semantically un-related can not be addressed by solely utilising formal semantics. In addition, for the disambiguation of specific tags, e.g., `beetle`, only some of the contexts should be taken into account e.g., {`beetle`, `car`}. If an explicit relation among the tags of a context exist, then this problem is eliminated. However, if the tags are related but the employed Knowledge Sources either do not cover the tags or do not declare their explicit relations, the co-occurrence frequency of the tags could be exploited to simulate their relatedness. Cattuto et. al [36] showed that statistically obtained relatedness matches the semantic measure in performance. Therefore, **alternative disambiguation measures that take into account statistical relatedness of tags should be considered** (see Chapter 7).

Finally, the query `fruit` caused lower precision because the algorithm exploited the direct hyponyms in the WordNet hierarchy of fruits, which were not compatible with the information need of the user. The user was looking for fruit images while the subsenses of fruit in WordNet include {`seed`, `pome`, `berry`, `achene`, `acorn`}. These are more abstract categorisations of fruit and do not cover well the tagspace. This is another case of a semantic entity (the WordNet synonym of fruit) that is conceptually

[9]The tags {`om1`, `multipleexposure`} are idiosyncratic, i.e., they do not refer to the concepts depicted in the image.

Figure 5.5: An image of a Volkswagen Beetle among plants

correct, but is not useful because its neighbours do not map well to the tagspace.

User Experience

To obtain an indicator of user satisfaction we assessed the user replies to questions Q3 to Q6. Their comments on comparatively using (S) and (T) are summarised in Table 5.2.

The users stated that the results returned by (S) were more relevant to what they had in mind and the presentation was more meaningful and helpful in finding what they were looking for. The structure of the results in (C) also generated more ideas for query reformulation. Additionally (S) returned more results and in cases that (T) did not return anything meaningful, (S) returned relevant results to the information needs of the user.

Pros	Cons
Returns more results	Does not categorise results in (B)
Meaningful presentation of results	Does not broaden the query keyword
Ideas for query reformulation	

Table 5.2: User Responses on regarding the use of (S)

Furthermore, they would prefer a sense-based grouping of results in (B) and they would like (C) to also give them broader results in the cases of low or no results. One such case is querying with `boletus` which is a type of mushroom and there were no resources tagged with it or its subordinates. In that case the users would prefer to see other types of mushrooms.

We note that the responses for Q5 demonstrate that in 21% of the queries on (S), there were incorrect results, while the same happens for the 12% of the cases in (T). This result on (T) indicated that searchers did not agree with the taggers. The users commented that they would not annotate this image with the query keyword and as a result, they consider it an incorrect hit for their search. This disagreement is inherent in folksonomies since users are free to tag the resources with any keyword. Finally, the increased value for (S) in Q5 was caused due to incorrect results such as the case of `beetle`.

5.4 Lessons Learnt

In the study presented in this chapter we compared the traditional search on flat folksonomies with a preliminary implementation of search on enriched tagspaces. We selected a dataset from Flickr related to plants and enriched it with the FLOR-1 enrichment algorithm. We implemented three search strategies in a web interface where users could ask keyword queries related to the domain of plants and had to compare their querying experience in using the baseline folksonomy approach (T) with the query

expansion facilitated by the enriched tagspaces (S). With the latter we obtained average normalised increase of 36% with average precision of 94%, which may have been influenced by the domain restriction of the dataset. The users were satisfied by the additional results from (S) and stated that their structured presentation generated more ideas for query reformulation. A qualitative investigation of individual user query keywords demonstrated cases where the Semantic Web could contribute further to the typical WordNet based approaches. For example, in the case of **beetle**, which was incorrectly mapped to insect from WordNet because the concept of car name is not included in WordNet.

Below we summarise the main outcomes of this study.

L5.1 The value of semantic entities with respect to the enrichment process depends not only on their richness on the Knowledge Sources of provenance, but also on how well their semantic neighbourhoods match the tagspaces.

L5.2 Statistical relatedness measures should be explored in order to exploit tag contexts which are not semantically related.

L5.3 WordNet allows satisfactory query expansion, therefore it should be considered as a Knowledge Source for enrichment.

L5.4 The presentation of results in groups is meaningful to the users and can help generate ideas for query reformulations.

Chapter 6

A Task Based Comparison of Online Ontologies and WordNet on Search

In this chapter we compare ontologies accessible via the Watson Semantic Web Gateway and WordNet as Knowledge Sources used for the purposes of search. We use them individually to extract two sense structures for a given tagspace and then exploit these two structures for search. We juxtapose the two sense spaces in terms of structure, we perform a user experiment in order to gain insights on their influence on search, and finally compare them against folksonomy based search.

6.1 Introduction

The work presented in Chapters 4 and 5 provided evidence on the value of WordNet in the enrichment of tagspaces and in search. WordNet is a long term, continuously maturing resource used for information retrieval, text classification and sense disambiguation and spans several domains. However, its evolution is relatively slow and often lags behind in the representation of novel terminology. At the same time it is a robust knowledge artefact of high quality. On the contrary, the ontologies on the web

originate from various sources and may encode more up-to-date knowledge compared to WordNet. For the same reason, though, they contain modelling and other errors and exhibit a high degree of heterogeneity.

In Chapter 4 we presented FLOR-1 and evaluated its enrichment precision and the tagspace coverage using a randomly selected sample. In that experiment we used WordNet to semantically expand the tags before matching them against semantic entities from ontologies. The experiment showed that the WordNet hierarchy was insufficient to determine whether two tags were semantically related (e.g., `lake` and `glacier`) therefore did not provide a good basis for disambiguation (L4.3). In Chapter 5 we experimented with FLOR-1 using a larger dataset from Flickr in order to obtain user insights on semantically-enabled search. Only the WordNet hierarchy was used to perform tag expansion and we observed that the results were satisfactory. This led to the hypothesis that WordNet could be used as a complementary Knowledge Source for the purposes of enrichment (L5.3). Aiming to address Research Question 2, which concerns the identification of alternative Knowledge Sources for tag enrichment, in this chapter we perform a comparative study of WordNet and ontologies available online in the context of folksonomy enrichment and search.

In the experiment we extract two sense structures from the two Knowledge Sources. For the purpose of assessing how they can address the issues of polysemy, synonymy and basic level variation we evaluate their structural properties using measures M3.2 - M3.5 (Section 3.6). Then, using the two sense structures we perform an experiment with semantically-enabled search that serves two purposes. First, in order to evaluate the semantic overlap with the tagspace (i.e., if the senses' neighbourhoods cover the tagspace, according to L5.1) we obtain the search results and apply the measure of mean normalised increase M3.11. The mean normalised increase depends on how well the sense expansion M3.6.2, which represents the sense's neighbourhood, maps to the tagspace. The second purpose of the search experiment is to compare the semantically-enabled search with standard folksonomy search in terms of result grouping and presentation.

6.2 Method

In this section we describe how we obtain two sense structures from WordNet and ontologies and describe the search mechanism used to query them.

6.2.1 Creation of Sense Spaces $\mathcal{S}_{\mathbf{KS}}$

We use two different strategies to enrich tagspaces with semantic structure. Strategy A uses WordNet and yields the sense structure $\mathcal{S}_{\mathrm{W}}$ and Strategy B uses online ontologies and produces $\mathcal{S}_{\mathrm{O}}$. Both $\mathcal{S}_{\mathrm{W}}$ and $\mathcal{S}_{\mathrm{O}}$ are structures similar to the one depicted in Figure 3.1. They are both built in two stages, common to both strategies.

Stage 1: First, the **potential meanings of a tag** are discovered by aligning it to appropriate senses. Strategy A relies on WordNet's synsets to find such senses, while in the case of Strategy B we employ a clustering mechanism which identifies a possible set of senses for a tag by combining information from multiple online ontologies. While in previous work (Chapter 4) we used disambiguation algorithms to precisely identify the meaning of a tag in a certain context, for the purposes of this comparative study we assign all possible senses to the generic tags. The reason is that we are interested in the richness and coverage of the Knowledge Sources over a tagspace and want to rule out any bias introduced by disambiguation methods. As a result, we assign all candidate senses to a tag, for the example of Figure 3.1 it would hold $senses(\texttt{apple}) = \{\mathrm{S}_2, \mathrm{S}_5\}$.

Stage 2: Second, we include **structural information among the senses** by reusing knowledge from the Knowledge Source. In particular, we consider all possible ancestors for each sense. For instance *Apple Inc.* is defined both as a *Company* and as an *Organisation*, $sup(S_4) = \{S_5, S_6\}$. In order to achieve a high degree of connectivity between the senses we consider the subsumption path up to the highest possible ancestor. We restrict this method to subsumption relations, as

these are present in both Knowledge Sources.

Strategy A: WordNet-Based Enrichment.

WordNet is a hierarchy of synsets each describing a sense. Most synsets are subsumed by at least one hypernym synset, subsume a set of hyponym synsets, and contain a set of words describing the same sense (synonyms).

For **sense selection**, we consider all the synsets that contain a given tag in their set of synonyms. Note that we consider only noun synsets as these have richer hierarchical information than other parts of speech. For each sense, we import in the structure $\mathcal{S}_W$ the corresponding synonyms of the sense. WordNet's matching mechanism automatically caters for lexical variations and plurals. To create a **structure of senses**, we import each sense's ancestor path up to the root of the WordNet hierarchy. Finally, for each sense we include the first level of hyponyms as subsenses.

Strategy B: Online-Ontology Based Enrichment.

In order to enrich the tagspace with $\mathcal{S}_O$, we explore online ontologies through the Watson[1] Semantic Web gateway. The sense selection is less straightforward in this strategy, because, unlike WordNet, the Semantic Web does not contain an established set of senses. To overcome this limitation, we use the clustering algorithm described in Section 4.4.1 which groups together entities that are sufficiently similar and therefore might denote the same sense. The process of **sense selection** from ontologies is as follows.

For each generic tag, we use Watson's API and we strictly match it against the id or label(s) of ontological concepts. For instance, `berry` is not matched against *Berry_Fruit* nor is `water` against *Water_Container*. This is done to reduce noise. By using multiple ontologies, the same concept may be defined more than once thus leading to different

[1] `http://watson.kmi.open.ac.uk`. The ontologies indexed in Watson during the experiment (May-June 2009) were approximately 9.000 and contained a total number of 460.000 classes (including redundancies).

types of redundancies. We use the entity merging methods described in Section 4.4.1 and the entity similarity measure M4.2.

For this experiment we set a low similarity threshold of 0.3 in order to achieve a maximum clustering result. In addition, we give the weights used in M4.2, W_G and W_L, a value of 0.5 in order to reflect the heterogeneity of online ontologies in terms of the richness of their lexical and structural information. For example, for the tag `banana` we obtained a single cluster of entities, because, according to our clustering algorithm there is only one sense of banana in all online ontologies. This cluster of entities contributes to the sense of *banana* with synonyms derived from the local names and labels {L_1: "*banana*", L_2:"*an elongated yellowish fruit which grows on palm trees*"}. L_2 was the label of one of the clustered entities. Different ontologists have different representation styles and may include a comment as a label. In addition, unlike in the case of WordNet, mapping of inflections is not covered by the Watson API's search mechanism and therefore lexical variations of the same concept will denote two different senses if they are not clustered by our algorithm. Issues such as lexical matching and entity redundancy need to be dealt with in Strategy B. All these are effects of the heterogeneity of ontologies.

To create the **structure of senses**, once the entity clustering is complete, for all the direct superclasses of the cluster's entities we iteratively get their superclasses up to the root of each ontology. For example, we obtain

Sense: ***Banana*** $\overset{subClassOf}{\longrightarrow}$ ***Fruit***

(Ontology 1) $\overset{subClassOf}{\longrightarrow}$ ***Tropical Fruit*** $\overset{subClassOf}{\longrightarrow}$ ***Fruit***

(Ontology 2) $\overset{subClassOf}{\longrightarrow}$ ***GroceryProduce***

(Ontology 3) $\overset{subClassOf}{\longrightarrow}$ ***Tree Fruit***

We notice that by adding this knowledge there is then one direct and one indirect relation between *Fruit* and *Banana*. We maintain as many subsumption relations as possible regardless of whether they are redundant, in order to support query expansion.

6.2.2 Query Mechanism

The query mechanism allows the exploration of the sense structures $\mathcal{S}_{\mathrm{KS}}$ created using the two Knowledge Sources. Algorithm 1 describes a querying process which first maps query keywords to appropriate senses, then retrieves the resources tagged with tags associated to these senses, and finally groups the result resources into meaningful groups, which are used as a basis for the presentation of the results.

Algorithm 1 Knowledge-based Querying

for all query keyword **k** **do**
  $\mathrm{g_k} = res(\mathbf{k})$
  $\mathcal{S}_{\mathrm{k}} = senses(\mathbf{k})$
  for all $S \in \mathcal{S}_{\mathrm{k}}$ **do**
    $\mathrm{g_k} = \mathrm{g_k} \cup \{\bigcup_{\mathrm{t} \in syn(S)} res(\mathrm{t})\}$
    for all $\acute{S} \in sub(S)$ **do**
      $\acute{\mathrm{g}}_{\mathrm{k}} = \bigcup_{\acute{\mathrm{t}} \in syn(\acute{S})} res(\acute{\mathrm{t}})$,
      for all r $\in \mathrm{g_k}$ **do**
        $Overlap(\mathrm{r}, \acute{S}) = \frac{|tags(\mathrm{r}) \cap syn(\acute{S})|}{|tags(\mathrm{r}) \cup syn(\acute{S})|}$
        if $Overlap(\mathrm{r}, \acute{S}) > MaxOverlap$ **then**
          move r to $\acute{\mathrm{g}}_{\mathrm{k}}$
        end if
      end for
    end for
  end for
  if $|\bigcup_{S \in \mathcal{S}_{\mathrm{k}}} sub(S)| < 4$ **then**
    for all $S \in \mathcal{S}_{\mathrm{k}}$ **do**
      for all $\dot{S} \in sup(S)$ **do**
        $\dot{\mathrm{g}}_{\mathrm{k}} = \bigcup_{\dot{\mathrm{t}} \in syn(\dot{S})} res(\dot{\mathrm{t}})$,
      end for
    end for
  end if
end for

The algorithm is based on the hypothesis that users are primarily interested in resources tagged with the exact query keywords, as well as with tags denoting more specific concepts. However, in cases where only a few resources are returned the user might also be interested in exploring resources tagged with more generic tags. For example, when searching for `fruit`, a user is likely to be looking also for resources annotated with the various types of fruit, such as `apple` or `tropical fruit`. Alternatively, if few

results are returned, it may be worth returning results associated with broader notions, such as **plant**. Accordingly, for a query keyword **k**, Algorithm 1 retrieves the relevant senses for **k** and creates a set of resources, $g_{\mathbf{k}}$, annotated with the synonyms of these senses[2] (Algorithm 1: 2-5). For each subsense of **k**'s senses, one group is created with the resources annotated with its synonyms (Algorithm 1: 6-7). Then all the resources tagged with **k** and its synonyms are compared against the synonyms of the subsense. The resource is moved to the group with whose sense it has the higher overlap. This is done in order to present the resources in specific groups (Algorithm 1: 8-11). For example, in a query for **animal**, items tagged with **animal** and **zebra** and items tagged with **zebra** are grouped together into a group created by the subclass of *Animal*, which is *Zebra*. If the number of subsenses is less than four, then the same process is repeated with the supersenses (Algorithm 1: 16-19). The threshold of four is selected because we further compare the knowledge-based querying with the cluster-based querying where the mean number of clusters per tag is 3.4 (Section 6.3).

6.3 Experiments

As a basis for our experiments, we used the MIRFLICKR-25000 [61] dataset proposed for ImageCLEF 2009. This contains 25000 images from Flickr with 69099 distinct tags. Although this is a dataset proposed for image analysis and 9% of the images are not tagged, the rest are tagged with a number of tags ranging from one to 75, spanning various domains.

We conducted three experiments. First, we enriched the dataset with strategies A and B and evaluated the enrichment in terms of quantitative and qualitative measures (Section 6.3.1). Second, we performed a user evaluation on search using the three systems built in Section 6.3.2. Finally, we used the user queries, $\mathcal{K}$, to quantify the mean normalised increase, M3.11, $\overline{|ninc(\mathcal{K}, \mathcal{S}_{\text{KS}})|}$ for each Knowledge Source (Section 6.3.3).

[2]Note that the synonyms of **k**'s senses are the synonyms of **k** including **k**.

6.3.1 Enrichment Evaluation

The values we obtained for the metrics defined for the evaluation enrichment in Section 3.6 are shown in Table 6.1. In terms of the tagset coverage of the two knowledge sources, we observe that WordNet covers more tags than online ontologies which is 26.3% of all tags, versus 16%.

	Measure	WordNet	Ontologies
M3.2	$\overline{\lvert syn(\mathcal{S}_{\text{KS}})\rvert}$	2.3	2.2
M3.3	$\overline{\lvert senses(\mathcal{T})\rvert}$	2.9	1.8
M3.4	$\overline{\lvert sub(\mathcal{S}_{\text{KS}})\rvert}$	2.7	1.5
M3.5	$\overline{\lvert sup(\mathcal{S}_{\text{KS}})\rvert}$	1.0	1.5
M3.11	$\overline{\lvert ninc(\mathcal{T},\mathcal{S}_{\text{KS}})\rvert}$	38%	39%

Table 6.1: Quantitative results of the enrichment evaluation

One of the reasons for the low lexical coverage by both Knowledge Sources is that, approximately 71.4% of the tags were not mapped to any of them. This was due to phenomena such as compound tag concatenation (**rowingboats**), misspellings (**rasberry**), non English tags (**chaminá**), idiosyncratic tags (**:D**), tags that are not defined in either source (**augor**) and phrases (**daughtersoftheamericanrevolution**).

Additionally, the major difference in coverage between WordNet and ontologies can be explained by the difficulty of anchoring tags to the concepts of these sources. We used strict matching to avoid the additional noise from ontologies, while WordNet has its own mechanism for matching of tags to synsets. In addition ontologies use different modelling styles to express the names of entities, using one or more of the following mechanisms: the local name(id), *rdf:label*, *rdf:comment* or even locally specified properties (for example *O2:name*). In addition, the delimitation of compound labels such as *zantedeschia__genus_zantedeschia*, *FloweringPlant* is inconsistent across ontologies.

In terms of the richness of the created structures, WordNet provides, on average, more

senses per tag (2.9) than ontologies (1.8). The amount of synonyms per senses are comparable in both sources, but important differences can be observed in the average number of more generic and more specific senses created in the two structures. Indeed, the structure created with WordNet, has a higher number of subsenses (2.7) on average than ontologies (1.5). Inversely, ontologies lead to more supersenses (1.5) than WordNet (1.0). This is because online ontologies often express different points of views, or cover more domains than WordNet does and therefore using the clustering mechanism produces more supersenses. For example, banana has four more generic senses in ontologies and only one in WordNet:

Synset: ***Banana*** $\overset{hyponym}{\rightarrow}$ ***Edible Fruit*** $\overset{hyponym}{\rightarrow}$ ***Fruit***

Sense: ***Banana*** $\overset{subClassOf}{\longrightarrow}$ ***Fruit***
(Ontology 1) $\overset{subClassOf}{\longrightarrow}$ ***Tropical Fruit*** $\overset{subClassOf}{\longrightarrow}$ ***Fruit***
(Ontology 2) $\overset{subClassOf}{\longrightarrow}$ ***GroceryProduce***
(Ontology 3) $\overset{subClassOf}{\longrightarrow}$ ***Tree Fruit***

We also observe variable hierarchical granularity between ontologies and WordNet. Indeed, as shown for *banana*, the WordNet definitions of terms tend to be more fine-grained than in ontologies. Additionally, differences in the granularity of the definitions can also be observed within WordNet itself. For example:

Synset: ***Orange*** $\overset{hyponym}{\rightarrow}$ ***Citrus*** $\overset{hyponym}{\rightarrow}$ ***Edible Fruit***

Synset: ***Apple*** $\overset{hyponym}{\rightarrow}$ ***Edible Fruit***

6.3.2 User-based Search Evaluation

To gain further insights on the impact of WordNet and ontologies on search, we created three search systems. Two systems implementing Algorithm 1 and querying $\mathcal{S}_{\mathrm{W}}$ and

$\mathcal{S}_O$ and one simulating baseline search in folksonomies. We then evaluated the three systems with the help of a group of users.

System Implementation

We developed two web interfaces which supported knowledge-based search as described in Algorithm 1. **System 2**, (S2) was based on the sense structure acquired from WordNet, $\mathcal{S}_W$ while **System 3**, (S3) exploited the structure created using online ontologies $\mathcal{S}_O$.

We also developed a web interface to simulate the cluster-based presentation of results currently provided by folksonomies, **System 1**, (S1)[3]. For this purpose we extracted the clusters of the query keyword using the Flickr API and for each cluster we created a group. Then we calculated the overlap of the tags of the resources tagged with the keyword with the tags of each cluster. The resource is then categorised under the group with which cluster it overlapped more.

All three web interfaces display the results grouped in meaningfully named groups. Figure 6.1 contains a screenshot with results from S1 for the query **sport** and Figure 6.2 the results for S3. For each group, there is a descriptive header which contains the title and the number of results per group. For S1 the title consists of the three most popular tags of the cluster (in accordance to the folksonomy clustering paradigm). For S2 and S3 the titles consist of the synonyms of the sense under which the results are clustered. For example, for the same query, **sport** S2 returned groups described as **track and field**, **skiing** or **judo**, while S3 had groups named **swimming**, **golf**, **football**, **hiking** and **stadium**. The "see all" link allows the user to view all the results of the group when there are more than five results per group.

[3] `http://www.flickr.com/photos/tags/TAG/clusters/`

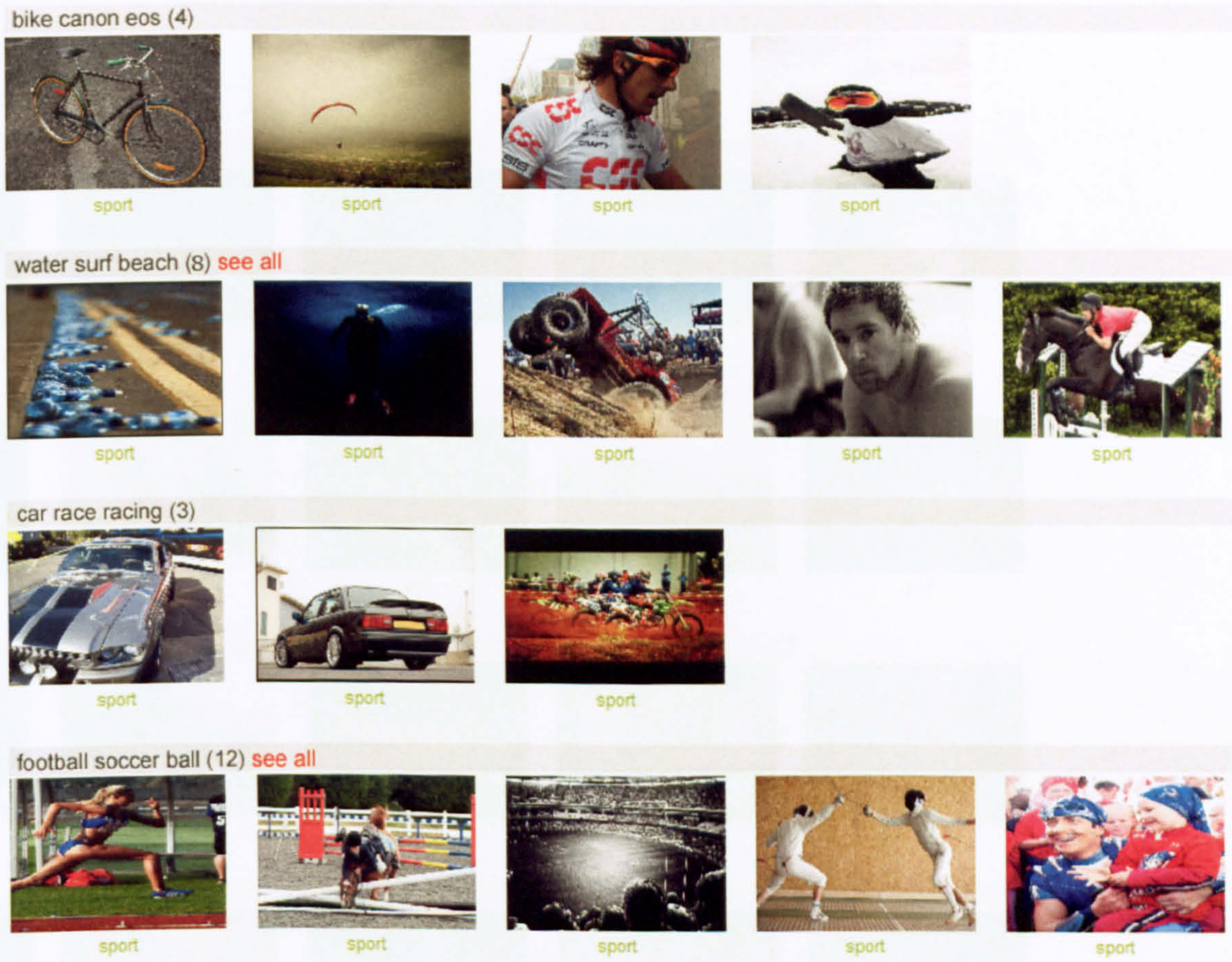


Figure 6.1: Result screenshot for the query *sport* in system S1.

User Experiment

In the second experiment we performed a user study. The user group consisted of 25 (expert and non expert) users with basic knowledge of image search on the web. Their task was to post at least three single keyword queries to systems S1, S2 and S3 without domain or any other restrictions. We limited the search to single keywords because we were interested in comparing the richness of the structures created in Step 1 per keyword. In addition we maintained the same terms to compare with S1, which simulates the cluster-based search which is only available for single keywords. We obtained 88 distinct queries and the evaluators had to report on their comparative experience on using S1, S2 and S3. More specifically, they had to report on the questions of Table 6.2. In Q2 and Q3 they had to select from a scale of 1 to 4; 1 being very

Figure 6.2: Result screenshot for the query *sport* in system S3.

unhelpful/all incorrect and 4 very helpful/all correct. They also had to report which results were most (ir)relevant/(in)correct and why for each query and system.

Question	S1	S2	S3
Q1: Did you find what you were looking for?	90%	85%	84%
Q2: How helpful was the presentation of the results and why?	2,9	2,8	2,8
Q3: Rate the number of correct versus incorrect results	3,3	2,8	3,1
Q4: Which is the best performing system?	35%	32%	33%

Table 6.2: User Experiment 2: Questions and Responses

Overall, S1 performs better than S3 which performs better than S2 as seen in Table 6.2. Considering that in S2 and S3 none of the results of S1 are excluded (Algorithm 1), a possible explanation for this result is the reported decrease in precision (Table 6.2, Q3). The users stated that **S1 performed better because there were less groups** and it was easier to navigate through the results. It should be stressed that all the results returned from S1 were tagged with the query keyword (see Section 6.3.2). S2 and S3 included results tagged with tags related to the keyword, thus increasing the number of groups. The number of groups in S3 was also increased by the **existence of overlapping senses with the same meaning** due to failure of the entity merging strategy. For example, in the case of car, two senses (one defining car as a *vehicle* and another defining car as *automobile*) were not merged, and appeared as two different senses in the results of S3.

The effect of irrelevant results was maximised by an additional factor. In some cases the users reported that the photos were tagged incorrectly. This can be further justified from the result of Q3:S1 = 3.3 (Table 6.2). An example of this is the query `tiger`. Among the groups containing photos of tigers, S1 returned a group headed with {`butterfly, shallowtail`} containing one image of a tiger butterfly. This was reported as incorrect because the user was unaware of this sense of the word `tiger` and no further explanation was given from the system. This is a common phenomenon arising from categorising photos based on clusters of tags derived from co-occurrence. The **relations among the tags are not clear** (e.g., *tiger butterfly is a type of butterfly*) and it is not possible to give a justification for the retrieval of results and their

categorisation in a particular group. This, however, would be possible if the knowledge *Tiger Butterfly* $\overset{subSense}{\longrightarrow}$ *Butterfly* was provided by a Knowledge Source.

In some cases the users reported that the presentation of S2 and S3 was more helpful even when the results returned by S1 were almost the same. According to them, the **images were presented under a meaningful category**. For example, for the query `horse`, S2 and S3 returned different groups for {`colt`, `palomino`} as opposed to the groups returned by S1 {`italy`, `cavallo`, `england`}. They found this distinction of results helpful for understanding the kind of horse depicted.

In the cases where the query keyword did not return meaningful results in S1, the users reported that S2 and S3 returned **a higher number and variety of results**. For example, querying for `soap`, most of S1's results depicted bubbles but S3 returned results depicting shampoo because *Shampoo* $\overset{subSense}{\longrightarrow}$ *Soap* was found in online ontologies. Equally, for `doggy` S1 retrieved only two images while S2 retrieved all images tagged with `dog` because `doggy` is one of the synonym terms for the sense of `dog` in WordNet. Finally the users were asked to select the system that performed better in all their queries (Q4). 35% of the users selected S1, 33% S3 and 32% selected S2. The responses to the rest of the questions of Table 6.2 justify this too. S1 performed better due to less groups of results. S2 and S3 returned better group descriptions and in addition S3 groups were judged to be more relevant than S2.

6.3.3 Quantitative Search Evaluation

Taking into account the user's query keywords, $\mathcal{K}$, and comments, we measured the approximate average normalised increase, (M3.11 in Section 3.6.2), $\overline{|ninc(\mathcal{K}, \mathcal{S}_{\mathrm{W}})|}$ for WordNet and $\overline{|ninc(\mathcal{K}, \mathcal{S}_{\mathrm{O}})|}$, for ontologies compared to the folksonomy search baseline. $ninc(\mathbf{k}, \mathcal{S}_{\mathrm{KS}})$ (M3.10 in Section 3.6.2) is the ratio of additional correct results returned by the expansion of $\mathbf{k}$ using the $\mathcal{S}_{\mathrm{KS}}$, divided by the total number of results as described in Section 3.6.2. Figure 6.3 shows the normalised increase for each keyword $\mathbf{k}$ from WordNet represented with dark lines and from ontologies' with light lines.

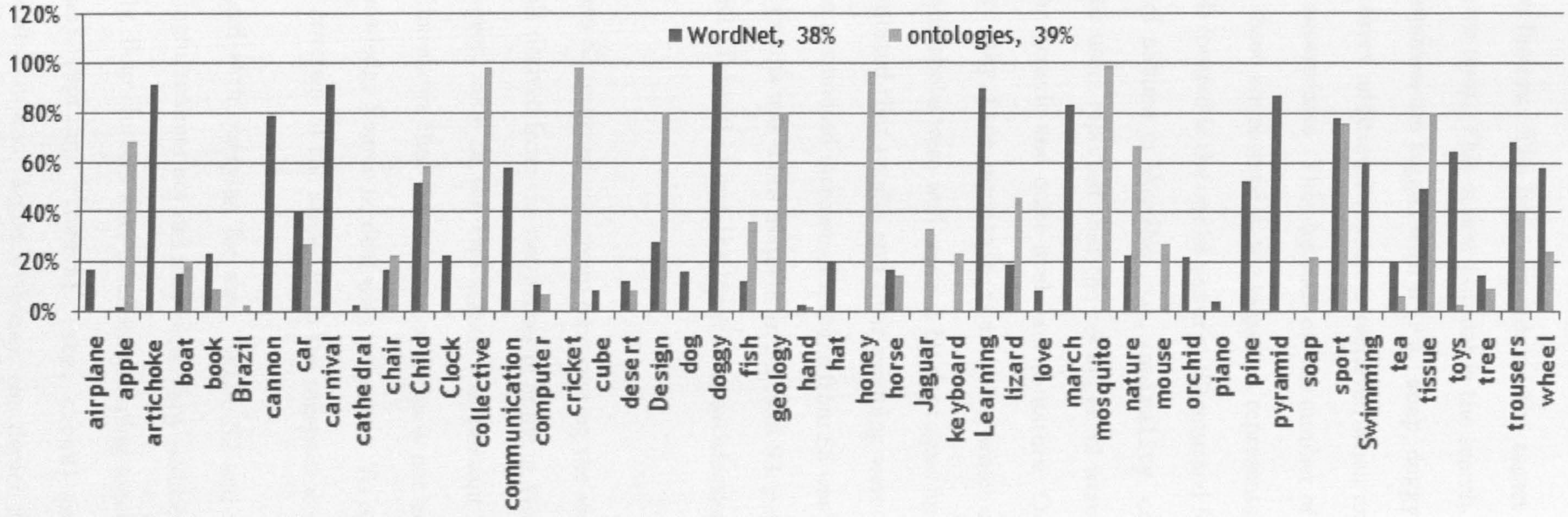


Figure 6.3: Normalised Increase in results on the user entered keywords

$ninc(\mathbf{k}, \mathcal{S}_{\mathrm{KS}})$ is affected by two factors. The first, depends on the tagset and is not relevant to the Knowledge Source used. This is how popular is the search keyword k_i in the tagset, i.e., how many resources are tagged with it (e.g., **soap**, **doggy**). The second factor, is the number of correct additional resources retrieved when expanding k_i with synonyms, subsenses and supersenses. This depends on the number of synonyms, sub/super-senses and how well these are covered in the tagset. A representative case of related senses that were not well covered is the one of **nature**. S2 returned four groups, one for each of the subsenses of **nature** in WordNet, i.e., {**animality**, **complexion**, **disposition**, **sociality**}. The users reported that the results of S2 were not significantly more than S1, they were generic and quite irrelevant to nature. On the other hand, S3 returned five groups for {**sky**, **fire**, **mountain**, **reef**, **rice**} which are ontological subclasses of **nature**. All the above were well covered by the tagset and, with the exception of **rice**, the users reported that results and their grouping were meaningful and satisfactory. As a result the normalised increase of **nature** from S3 was significant. The phenomenon of irrelevant groups was more frequent in S2 than S3, justifying the lower normalised increase for S2 (Table 6.1) and the lower user satisfaction with this particular system (Table 6.2).

When querying for **apple**, system S3 returned one group representing the sense of fruit. S2 returned two groups for apple derived from the two senses of apple in WordNet, i.e., **fruit** and **fruit tree**. However, none of the two senses was relevant to the the sense of **computer company**. This shows that the number of senses is not necessary an indicator of the ability of a Knowledge Source to deal with polysemy. To reach such a conclusion, information on the coverage of the senses from the tagspace is required.

Another useful outcome emerged with querying for **may**. While S2 and S3 did not present the results in any meaningful manner nor did they return any additional results, S1 performed quite satisfactorily. Four clusters were returned grouping together images tagged with {**england**, **london**}, {**spring**, **flowers**}, {**sky**, **cloud**} and {**paris**, **france**}. A plethora of photos shot in and tagged with **may**, can depict flowers, sky,

cities and so on but may depict nothing that can symbolise the month May. This is a type of idiosyncratic tagging and no Knowledge Source can supply formal relations between `may` and these tags since there are no formal relations among them. Nevertheless, for this type of **idiosyncratic tagging the clustering of results based on frequent tag co-occurrence is quite efficient**.

6.4 Lessons Learnt

In this study we explored how formal knowledge sources, WordNet and online ontologies, can improve folksonomy search and which of them performs better. We evaluated them qualitatively and quantitatively in terms of tagspace enrichment and user satisfaction comparing the knowledge-based search to cluster-based search.

In terms of tagspace enrichment, WordNet outperformed ontologies in most of the measures. It provided more senses per tag and more synonyms per sense than ontologies and lexically covered a higher percentage of tags than ontologies. WordNet and ontologies returned comparable measures for subsenses and supersenses. The above measures indicate that WordNet performed better in terms of sense richness and similarly to ontologies in terms of structure. However, in the user evaluation the ontologically created structure performed better in search than the WordNet created structure. The expansion of the query keyword with terms from ontologies returned a higher number of results compared to the expansion provided from WordNet despite the fact that WordNet provided a richer structure. This indicates that ontological structures of senses map better to the tagspace.

Comparing the knowledge-based search to folksonomy search, indicated that **users prefer the number of groups to be concise**, similar to cluster-based search but the **explanation of the results to be more intuitive** similar to knowledge-based search. In addition, search problems caused by idiosyncratic tagging can be addressed

better by statistical methods rather than formal knowledge sources.

In Chapter 7 we show how information from folksonomies, ontologies and WordNet can be combined to achieve better sense discovery for tags. In particular, we use hybrid disambiguation techniques in order to assign the tags to the most relevant senses using both knowledge sources. In addition, we extend strategies A and B so that they exploit more entities and relations from each Knowledge Source and use a more elaborate entity merging mechanism.

In the following we summarise the key outcomes of the experiment presented in this chapter.

L6.1 WordNet provides more synonyms for a sense compared to ontologies but neighbourhoods of senses derived from ontologies map better to the tagspace. Therefore a combination of the two Knowledge Sources would be beneficial for the enrichment of tagspaces.

L6.2 Statistically clustering the results returns less groups, caters for idiosyncratic tags but does not explain why a result belongs to a group.

L6.3 Semantically-enabled search returns more meaningfully organised results but the number of groups should be restricted.

L6.4 The existence of senses with the same meaning has an adverse impact on search.

Chapter 7

Improved Version of Folksonomy Enrichment Algorithm

In this chapter we describe the improved version of the FLOR enrichment algorithm. We describe how the new version is influenced from the outcomes of the studies performed using the previous version and detail the individual steps and processes.

7.1 Introduction

In this chapter we describe the final version of the enrichment algorithm, FLOR-2, based on the requirements that emerged from the analysis we conducted in Chapters 4 to 6. The goal of the algorithm is, given a tagspace $\mathcal{T}$, to create a semantic structure that contains the meaning of the tags in $\mathcal{T}$ and their relations. The desired output is the structure demonstrated in Section 3.5, Figure 3.8. In this structure, each tag $\mathtt{t}$ is connected to a sense S that describes its meaning and each sense is connected with other senses in the structure. Finally, each sense is linked to the semantic entity(ies), from which it originates. The production of this output is dependent on the discovery of appropriate semantic entities, the creation of suitable senses, the disambiguation of

tag meaning and the discovery of relations among the senses.

7.2 FLOR-2 Overview

The implementation of FLOR-2 was guided by the outcomes of the studies described in chapters 4 to 6. In these studies we performed two types of analysis. The first aimed at evaluating the output of FLOR-1 in terms of enrichment. We explored the coverage of tags in the employed Knowledge Sources and the correctness of assigning semantic entities to tags. The second analysis assessed the value of the enriched tagspaces from the perspective of search. Below we explain how the outcomes of this analysis translate to design requirements for the final version of FLOR.

7.2.1 Design Requirements

The first outcome of our previous investigation regarded the inclusion of WordNet in the enrichment procedure. In our first studies (Chapters 4, 5) we used WordNet as a thesaurus to disambiguate the meaning of tags and expand them with synonyms and hypernyms. We did that based on the assumption that hierarchical relations may hold between the tags of a tagset and used a hierarchical similarity measure for their disambiguation. Experimenting with this approach showed that the implicit relations among the tags of a tagset are usually not hierarchical. As a result, similarity measures based on subsumption are not adequate for disambiguating tags in such contexts (L4.3) and statistical correlations between tags may need to be considered to improve disambiguation (L5.2).

Rq_1: Disambiguation should exploit statistical relatedness in addition to formal knowledge.

The second outcome of our experiments was the low coverage of tags against the Knowledge Sources. Our study presented in Chapter 6 where we did a search-based comparison of WordNet and ontologies, showed that their value in folksonomy search and enrichment is similar, while WordNet's performance on search was satisfactory (Chapter 5). As a result we decided to use WordNet as a Knowledge Source for entity discovery rather than as a thesaurus for expansion (L5.3, L6.1).

Rq_2: Entity Discovery should exploit WordNet as a Knowledge Source for entity discovery rather than as a source for disambiguation and semantic expansion.

The search experiments presented in Chapter 6 showed that the existence of overlapping senses (senses that have the same meaning but have not been merged into one) has an adverse impact on search (L6.4). As a result we revise the entity merging strategy in an effort to integrate all senses that convey the same meaning.

Rq_3: Sense Discovery should integrate all sufficiently similar senses.

Rq_3 can provide additional value to the semantic aggregation phase. By clustering together senses with heterogeneous neighbourhood we are bound to achieve a higher connectivity of the final structure.

7.2.2 Data Structures and Components

The algorithm is visualised in Figure 7.1 and involves the following data structures:

Tag and Tagset represent a specific tag and the tagset in which it occurs. The tag is represented by a disk with the label `t` and the tagset with the oval labeled with `T`. The tagsets are the input of FLOR-2 and are handled by the lexical processing

phase (Figure 7.1: steps 1 and 2). They are also the input for the sense disambiguation step. While sense discovery takes as input individual tags and returns as output a set of candidate senses to the sense repository, tag disambiguation requires as input the whole tagset because it expresses the context based on which the correct senses will be selected for its tags.

Semantic Entity is represented by the polygon labeled with *e*. The semantic entity (Section 3.4) is extracted from the Knowledge Sources and is the output of the entity discovery steps (Figure 7.1: steps 3 and 10). The semantic entities are then subject to entity filtering (Figure 7.1: steps 4 and 11) and the output of this phase is passed as input to the sense creation step.

Sense is introduced as the output of sense creation (Figure 7.1: steps 5 and 12) and is designated with the diamond labeled with *S*. Senses are added to the sense repository where they remain during the lifecycle of the algorithm. They are initially linked to tags as **candidate senses** by the sense discovery step. After the sense disambiguation (Figure 7.1: step 8) the relation between tag `t` and sense *S* is made explicit (see Chapter 3, *hasDefinition*) and is added to the output. The sense *S* that explicitly defines the meaning of the specific tag `R_t` is called **assigned sense** to `R_t`.

Sense Relation is the last data structure created by FLOR-2 and is designated by a pair of senses. Sense relations are added to the output during relation definition (Figure 7.1: step 9) and structure integration (Figure 7.1: step 14).

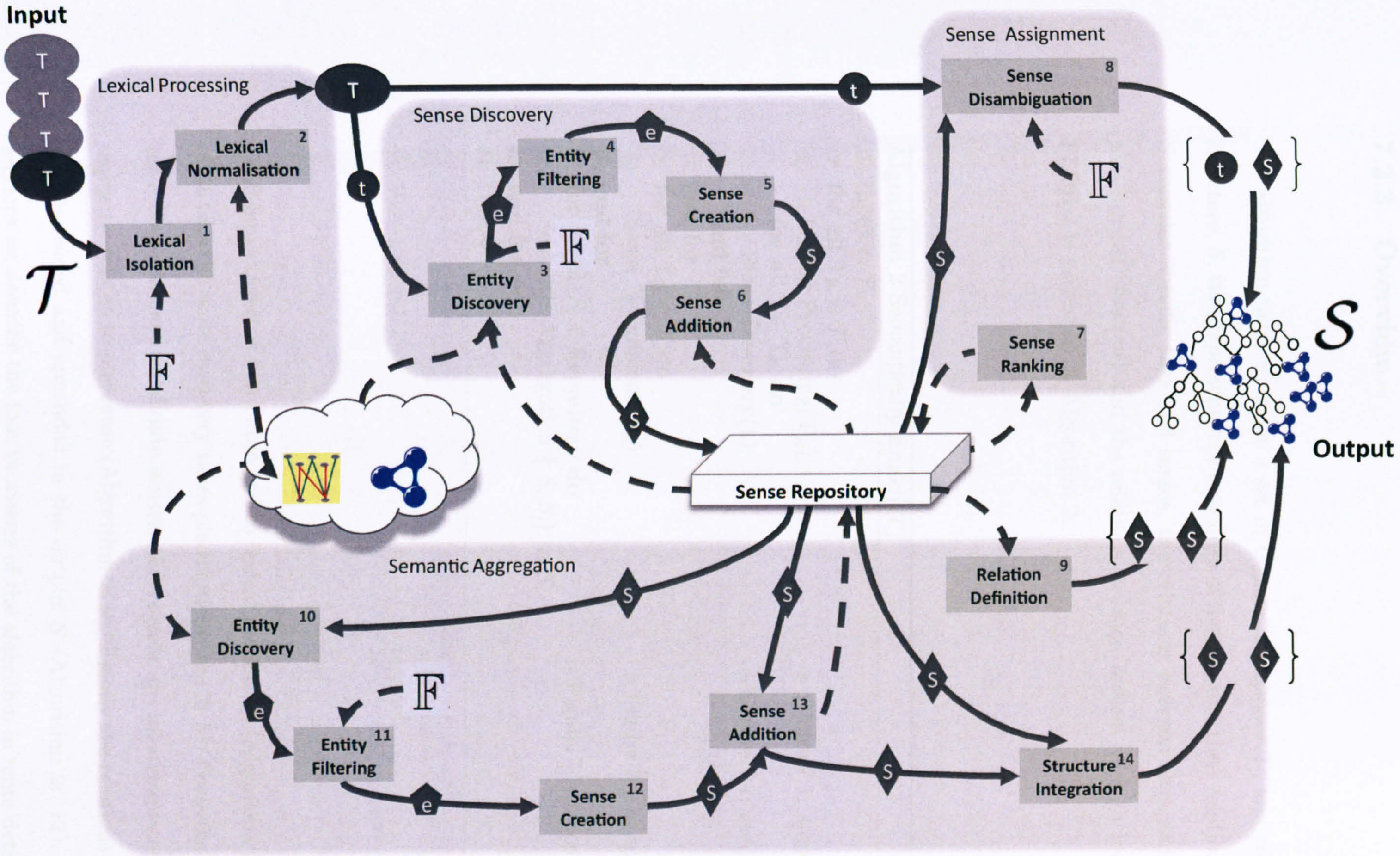
Figure 7.1 also depicts the components and resources used by the algorithm:

Knowledge Sources (defined in Section 3.4) are the online ontologies indexed by the Watson repository and WordNet. Knowledge Sources are queried during the entity discovery step (Figure 7.1: steps 3 and 10) when trying to locate appropriate semantic entities.

Sense Repository is a registry of all the senses that are encountered during the lifecycle of the algorithm. It contains all the candidate senses of the tags, the assigned senses for each tag and other useful information (required for sense ranking, sense addition, and relation definition). The sense repository is a central component and is used by the majority of the processes of FLOR-2. When the transaction of a process with the repository yields a specific object (e.g., a sense S) we use a solid line in Figure 7.1 annotated with the object type (e.g., a diamond labeled with S). Alternatively, when the repository is queried for generic information[1] the transaction is designated with a spaced line.

Folksonomy sources represent the components which provide the algorithm with statistical information about the tags. In Figure 7.1 they are exemplified with the symbol $\mathbb{F}$. They provide information about the clusters of tags (Section 3.2), which represent the statistical distribution of tags over a tagspace. FLOR-2 exploits existing clusters as the implementation of methods for tag clustering is out of the scope of this work. Several clustering methods exist in the literature (Chapter 2) and some folksonomies already provide clusters of frequently co-occurring tags (e.g., Flickr, Delicious).

[1]For example, step 3 checks the repository for already encountered candidate senses for a tag, prior to querying the Knowledge Sources. If such senses exist, no further steps are taken for the tag within the sense discovery step.


Input
T
T
T
T
Lexical Processing
Lexical Isolation 1
Lexical Normalisation 2
Sense Discovery
Entity Discovery 3
Entity Filtering 4
Sense Creation 5
Sense Addition 6
Sense Assignment
Sense Disambiguation 8
Sense Ranking 7
Sense Repository
Semantic Aggregation
Entity Discovery 10
Entity Filtering 11
Sense Creation 12
Sense Addition 13
Relation Definition 9
Structure Integration 14
Output

7.2.3 Overview

The algorithm takes as input a set of tagsets, a tagspace $\mathcal{T}$, and returns a semantic structure, $\mathcal{S}$, that represents the meanings of the tags and their relations. The structure $\mathcal{S}$ contains specific tags and senses, the relations between tags and their associated senses (*hasDefinition*) and the relations between senses. A high level description of FLOR-2 is presented in Algorithm 2.

Algorithm 2 Semantically Enrich($\mathcal{T}$)

```
1: Input = T
2: for all T_R ∈ T do
3:     Lexical Processing(T_R)          ▷ Figure 7.1: steps 1-2, Section 7.3
4:     for all t ∈ T_R do
5:         Sense Discovery(t)           ▷ Figure 7.1: steps 3-6, Section 7.4
6:     end for
7: end for
8: for all T_R ∈ T do
9:     Sense Assignment(T_R)            ▷ Figure 7.1: steps 7-8, Section 7.5
10: end for
11: for all S_i, S_j ∈ Repository do     ▷ Figure 7.1: steps 9-14, Section 7.6
12:     Semantic Aggregation { S_i,S_j} to S
13: end for
14: Output = S
```

Each input tagset is first subjected to lexical processing (Algorithm 2: 3) and for each of its tags the sense discovery takes place (Algorithm 2: 5). Once the tags of all tagsets have been assigned candidate senses, the tagsets are disambiguated and each tag is explicitly related to one sense (Algorithm 2: 9). Finally the relations among the senses are discovered and appended in the output $\mathcal{S}$ (Algorithm 2: 12). In the following sections we describe the four processes of the algorithm in more detail.

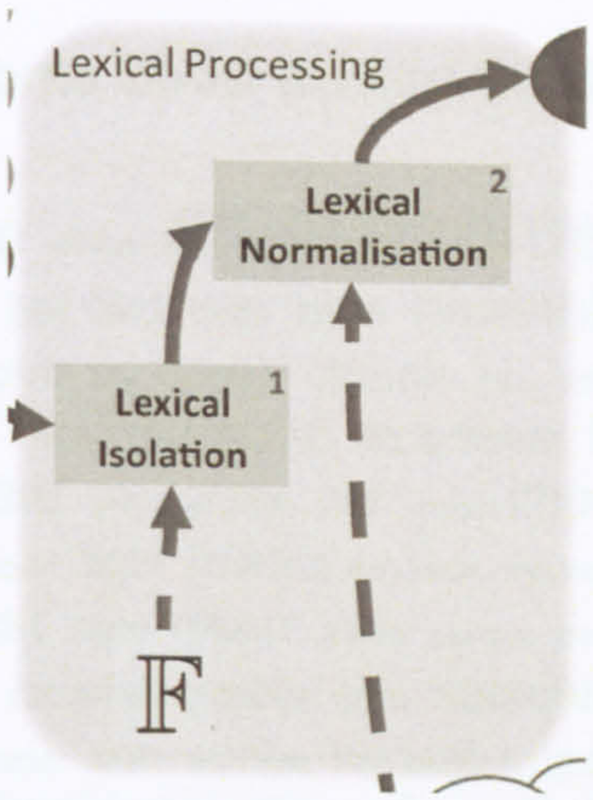


Figure 7.2: The Lexical Processing Phase

7.3 Lexical Processing

The Lexical Processing phase serves two purposes. First, to isolate tags that cannot be further enriched and would slow down the overall process. Second, to lexically enrich the tags so that we can achieve a high coverage of tags with respect to the Knowledge Sources.

7.3.1 Lexical Isolation

The isolation of tags has been previously discussed in Section 4.2.1. To the best of our knowledge, the only work that automatically identifies tag types (based on the analysis of Golder and Huberman [51]) is presented by Takeharu et.al in [44]. In this work, the authors identify subjective tags (i.e., usually expressing user opinions on the resource, rather than describing the resource content) which they exclude from their further analysis. The other approaches discussed in Chapter 2 employ various heuristics for the isolation of less useful tags. Based on the outcomes of our previous experiments and using a variety of heuristics we identified the types of tags that have a low value for the enrichment process. These types of tags are described below.

Figure 7.3: An example of Idiosyncratic Tagging in Flickr

Idiosyncratic tags

Idiosyncratic tags represent tags used for personal reference, opinion expression and participation in a social group. An example of idiosyncratic tagging from Flickr is presented in Figure 7.3. This is the tag cloud of the most popular tags of a group named "*An Awesome Shot! Invited Photos ONLY!!*"[2]. Such are social groups created by users with shared interests. The users are required to tag the photos they submit in the group pool with abbreviations of the group name. In Figure 7.3 we note that the tag `anawesomeshot` is the most popular in the group followed by other idiosyncratic tags expressing other types of group idiosyncrasies for example, {`abigfave`, `naturesfinest`, `thebestphotographer`}.

Such tags do not represent the concepts depicted in the resource, as a result their value to the semantic structure is low. We identify these tags using the following heuristic. For each generic tag `t`, we identify all the resources it tags: $R = res(\texttt{t})$. $R_G \subseteq R$ is the set of resources that belong to the pools of such groups. We use the ratio $\frac{R_G}{R}$ to decide if the tag is idiosyncratic or not. If the ratio is closer to 1, this means that most of

[2] 51,034 members and 385,801 photographs on August 2010

$\mathcal{T}es(\mathtt{t})$ belong to such groups, therefore it is likely that `t` is idiosyncratic.

Low frequency tags

Low frequency tags are used by few users and may include some types of idiosyncratic tags. These may express personal opinion e.g., `horriblywrong` or describe a vague or non-commonly understood concept, e.g., `mariasbirthday2009`. To identify these types of tags we use statistical information. For example, if they do not belong to clusters of frequently co-occurring tags, it means that they are not frequent.

Lexically-noisy tags

These are tags that contain special characters, such as `:D`, numbers, `top111`, and phrases `daughtersoftheamericanrevolution`. The identification and exclusion of these tags is straightforward and is performed using string length and character filters.

Depending on the nature of the input tagspace there is a possibility to include or exclude different types of tags by implementing additional isolation methods. The lexical isolation is the first step in the process of FLOR-2 (Figure 7.1) and requires additional information from $\mathbb{F}$ such as, the number of clusters a tag belongs to (for the infrequent tags) and the groups that a resource belongs to (for the idiosyncratic tags).

7.3.2 Lexical Normalisation

The lexical normalisation (step 2) is essential in order to achieve a better anchoring of tags to the various Knowledge Sources. For example the tag `flowers` is normalised to `flower` which is also used for querying in Knowledge Sources. In addition, in cases of tags such as `santaBarbara`, the different delimitations such as `santa barbara`, `santa-barbara` will also be included in the inflections of the tag in order to maximise

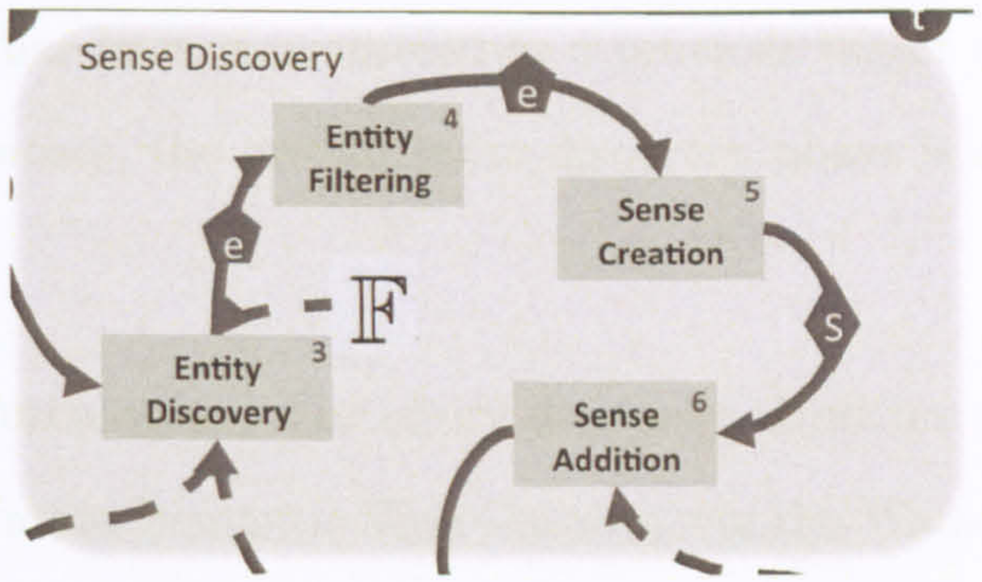


Figure 7.4: The Sense Discovery Phase

the coverage. To normalise the tags we use a combination of string processing and lexical information derived from WordNet (such as the singular inflections of a word, e.g., "flowers" maps to "flower").

7.4 Sense Discovery

The role of sense discovery is essential to the overall performance of FLOR-2 as requirements Rq_1, Rq_2 and Rq_3 are all addressed in various steps of this phase. In the following sections we describe the steps of sense discovery and how they address these requirements.

7.4.1 Entity Discovery

In the first instance, entity discovery (Figure 7.1: step 3) takes as input a tag `t` and queries the Knowledge Sources for entities that possibly match it (using the processes described below). Due to the high probability of a tag appearing in more than one tagset, the sense repository maintains all candidate senses for a tag `t`. These are all the senses S for which `t` belongs to their synonym set, i.e., $\texttt{t} \in syn(S)$. Therefore, prior to searching for new semantic entities the entity discovery step checks the repository. If no candidate senses are found in the repository, the entity discovery proceeds in

the manner described below. If `t` has been encountered in a previous tagset and has already been assigned candidate senses, the overall sense discovery phase is omitted for `t`.

Entity Discovery from online ontologies. The entity discovery algorithm exploits all the ontologies indexed in the Watson Semantic Web Gateway via the Watson API. Each tag `t` is used in the search mechanism of the Watson API in order to locate appropriate semantic entities. As appropriate entities we consider the classes and individuals (instances), which contain `t` in their lexical representation. The lexical representation of a semantic entity consists of its labels (denoted with "rdf:label" [34]) and its local name (ID). Due to the high heterogeneity of the modelling styles among the online ontologies (an outcome of the study presented in Chapter 6), various heuristics were employed in order to identify other possible lexical representations and increase tagspace coverage. For example, we also retrieve entities that contain the tag in their "rdf:comment" literals when the length of the comment is not longer than two words. The anchoring of tags to semantic entities is initially performed using strict matching, however, in case no results are returned we use flexible matching (e.g., **`berry`** is matched against the delimited **`berry_fruit`**, but **`tea`** is not matched against **`teacher`**).

Entity Discovery in WordNet. To satisfy Rq_2, WordNet is exploited as a Knowledge Source for the discovery of semantic entities. In this case the process of discovering semantic entities is more straightforward. The semantic entities are all the WordNet noun synsets which contain the tag in their set of synonyms. We use only nouns because they are hierarchically related to each other.

The output of the entity discovery step for a tag is a set of semantic entities, which are then subjected to entity filtering.

7.4.2 Entity Filtering

The need to perform filtering (Figure 7.1: steps 4 and 11) of the semantic entities emerges from the following characteristic of the knowledge indexed in Watson. The quality and richness of this heterogeneous knowledge is variable. While there is a plethora of useful ontological entities which can be efficiently reused, there are also entities which do not contribute to the enrichment process. Their enrichment value is low and thus are filtered out of the process. Such entities are:

Structurally poor entities. These entities do not have any relations to any other semantic entities nor lexical representations (apart from the label that was used to retrieve them). As a result, they are not useful to the creation of senses or the creation of an interconnected semantic layer. Such entities mainly originate from online ontologies.

Entities originating from Semantic Documents with low enrichment value. Watson indexes all web documents that contain semantic descriptions. In many of the cases, these documents are automatically generated descriptions of news feeds or user blogs. The majority of their entities are individuals and contain information irrelevant to the enrichment process. For example, in a search for `cat`, we obtain entities such as:

Individual: http://.../my/cat $\xrightarrow{title}$ `My 9 rules, Cat's Profile`
$\xrightarrow{generatorAgent}$ **http://www.talkdigger.com**
$\xrightarrow{type}$ **http://xmlns.com/foaf/0.1/Document**

We distinguish these documents by matching their domain names against a set of "stop-domains"[3]. In addition, we measure the number of classes such semantic documents define. We rule out the ones that contain less than 3 classes. This is an indication that such semantic documents contain a low number of semantic relations and thus their contribution to the semantic enrichment is bound to be low.

[3] According to the stop-words paradigm.

Entities with folksonomically-low value. Certain entities from online ontologies are defined in a manner that does not contribute to the enrichment process because they can not relate to other tags in the tagspace (L5.1). For example:

Class: **http://.../small3#food** $\xrightarrow{label}$ `food` $\xrightarrow{subClassOf}$ **http://...l3#DEFAULT_ROOT_CONCEPT**

Such an entity may be valuable in the context and for the purposes of the use case it was created for but not for a generic case such as the enrichment of folksonomy tagspaces. Such entities are ruled out by exploring the folksonomic value of their semantic neighbourhood. The folksonomic value is measured by folksonomy resources tagged with labels derived from the entities' semantic neighbourhoods. For example, combinations of `default`, `root` and `concept` are not related to any resource and as a result this entity is judged to be of low value and is filtered out of the enrichment process.

The process of entity filtering is required in order to overcome the above phenomena introduced by the paradigm of existing knowledge reuse. Similar issues on knowledge reuse have been highlighted by Lopez et.al in [77]. Unfortunately, no formal methods have been proposed for the task-based evaluation, selection and reuse of ontological knowledge. Therefore, in the scope of this work we address the issue of reuse by employing heuristics tailored to this approach. The output of the entity filtering is a set of semantic entities that do not exhibit the phenomena described above.

7.4.3 Sense Creation

The entities that qualify through the entity filtering are used to create senses according to Definition 6 of Section 3.4 (and the schema of Figure 3.7). The transformation of semantic entities to senses is necessary in order to achieve optimal sense integration

Figure 7.5: An example of a semantic entity returned for *Lake*

according to Rq_3. In the following sections we demonstrate how senses with different origin (from different ontologies and WordNet) are compared against each other and, with the help of a sense similarity function, are integrated into one.

Sense creation (Figure 7.1: steps 5 and 12) converts the semantic entities to senses. In Section 3.4 we exemplified the creation of senses from one ontological entity and one synonym for `apple`. We showed how the information of the entity is transformed to the respective attribute of a sense. For example, the synonyms of synsets and the lexical information of ontological entities (labels and local names) create the set of synonyms for a sense. Figure 7.5 depicts one ontological entity retrieved from Watson in search for `lake`. The sense created from this entity is:

$$\textbf{Sense:}\ \textit{lake} \xrightarrow{subSenseOf} \textbf{\textit{http://a.com/ontology\#WaterBasin}}$$
$$\text{PilotPolution1} \xrightarrow{superSenseOf} \textbf{\textit{http://a.com/ontology\#Pond}}$$
$$\xrightarrow{isFoundIn} \textbf{\textit{http://a.com/ontology\#Lake}}$$
$$\textit{“the exact difference between a lake [...] distinction.”}$$

In this sense of lake the following points should be considered.

- There is only one synonym for the sense (the local name of its semantic entity of origin, since no other appropriate lexical descriptions exist) and there is a natural language description which is transformed to the gloss of the sense. The value of the relation ***isFoundIn*** is the semantic entity, which is the **origin** of this sense. In the FLOR ontology (Figure 3.7) we specified that the relation *isFoundIn* has cardinality "one-to-many", which means that one sense can be found in many semantic entities. In the following sections we demonstrate how the integration of senses yields this result. If the semantic entity does not have an ontologically specified unique identifier (URI), for example it is a WordNet synset, we create one dereferenceable URI which consists of its offset[4].

- We also observe the exclusion of the disjointness relations from the information of the sense. These relations are meaningful in an ontological context and the only type of WordNet relations that could be equivalent to these are the antonymy relations. Yet, our approach does not exploit relations which imply negation.

- We also note that the new sense of lake is related to the neighbour entities of the original semantic entity (rather than being related to a sense). This is a temporary state, which will be addressed by the phase of semantic aggregation. There, the relations that hold between the sense and the neighbourhood of the original entity will be leveraged to relations between senses.

- Finally, the relation between the sense *Lake* and the original entity's superclass, *WaterBasin* is *subSenseOf* rather than *rdfs:subClassOf*. The latter is valid only among classes while the senses created by FLOR-2 are instances of the class *flor:sense*. As a result *rdfs:subClassOf* is not appropriate to define subsumption among them.

[4]An offset is a WordNet-specified unique identifier for synsets.

Consider two entities a and b that are transformed into senses A and B. The relations of a and b are leveraged to relations between A and B as follows:

1. Subordinate
   - $a \xrightarrow{rdfs:subClassOf} b \Longrightarrow A \xrightarrow{flor:subSenseOf} B$
   - $a \xrightarrow{wn:hyponym} b \Longrightarrow A \xrightarrow{flor:subSenseOf} B$
   - $a \xrightarrow{rdf:type} b \Longrightarrow A \xrightarrow{flor:instanceOf} B$
   - $a \xrightarrow{wn:instance} b \Longrightarrow A \xrightarrow{flor:instanceOf} B$
2. Superordinate
   - $b \xrightarrow{rdfs:subClassOf} a \Longrightarrow A \xrightarrow{flor:superSenseOf} B$
   - $a \xrightarrow{wn:hypernym} b \Longrightarrow A \xrightarrow{flor:superSenseOf} B$
   - $b \xrightarrow{rdf:type} a \Longrightarrow A \xrightarrow{flor:hasInstance} B$
   - $a \xrightarrow{wn:hasinstance} b \Longrightarrow A \xrightarrow{flor:hasInstance} B$
3. Meronymy
   - $a \xrightarrow{wn:meronym} b \Longrightarrow A \xrightarrow{flor:isPartOf} B$
   - $a \xrightarrow{wn:holonym} b \Longrightarrow A \xrightarrow{flor:hasPart} B$

   Although the subsumption relations in ontologies are universally stated with *rdf:type* and *rdfs:subClassOf*, the meronymy relations are only explicit in WordNet. These include all types of meronymy (membership, substance, part).

In the next section we describe how the created senses are compared against existing senses and are maintained in the sense repository.

7.4.4 Sense Addition

The sense addition step (Figure 7.1: steps 6 and 13) takes as input the senses produced by the sense creation step and adds them to the sense repository. The process of sense addition is depicted in Figure 7.6. Each new sense is compared against the relevant

senses that exist in the repository. This is important for the satisfaction of Rq_3 which requires optimal sense integration in order to avoid the redundancy of senses that describe the same meaning. Under-merging the senses would lead to the existence of more representations of the same meaning in the tagspace and would cause problems in the exploitation (Chapter 6) of the sense structure. On the other hand, over-merging the senses would lead to senses that may convey more than one meaning and this would be a conceptual error.

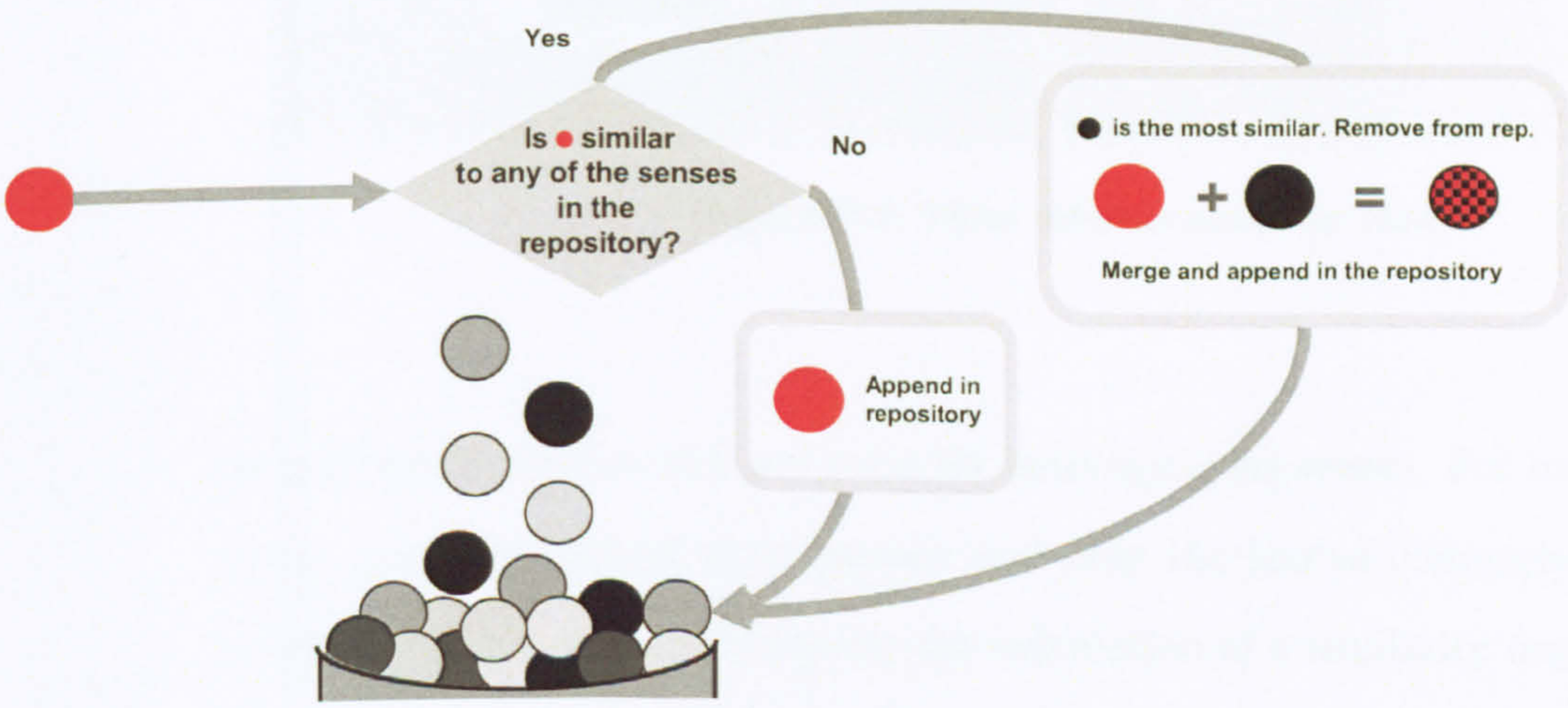


Figure 7.6: Adding a new sense in the Sense Repository

Sense comparison is done based on the sense similarity measure described below. If the similarity is low then the new sense is added to the repository. If the similarity is high then the existing sense with which this high similarity is achieved is merged with the new sense. The integration of the senses involves the creation of a new sense with all the properties of its original senses (see the detailed process below).

Sense Similarity

In Chapter 4 (Section 4.4.1) we presented a method for clustering ontological entities based on the similarity measure M4.2. We also used this approach in Chapter 6 where the following issues emerged. In some cases similar senses were not merged because

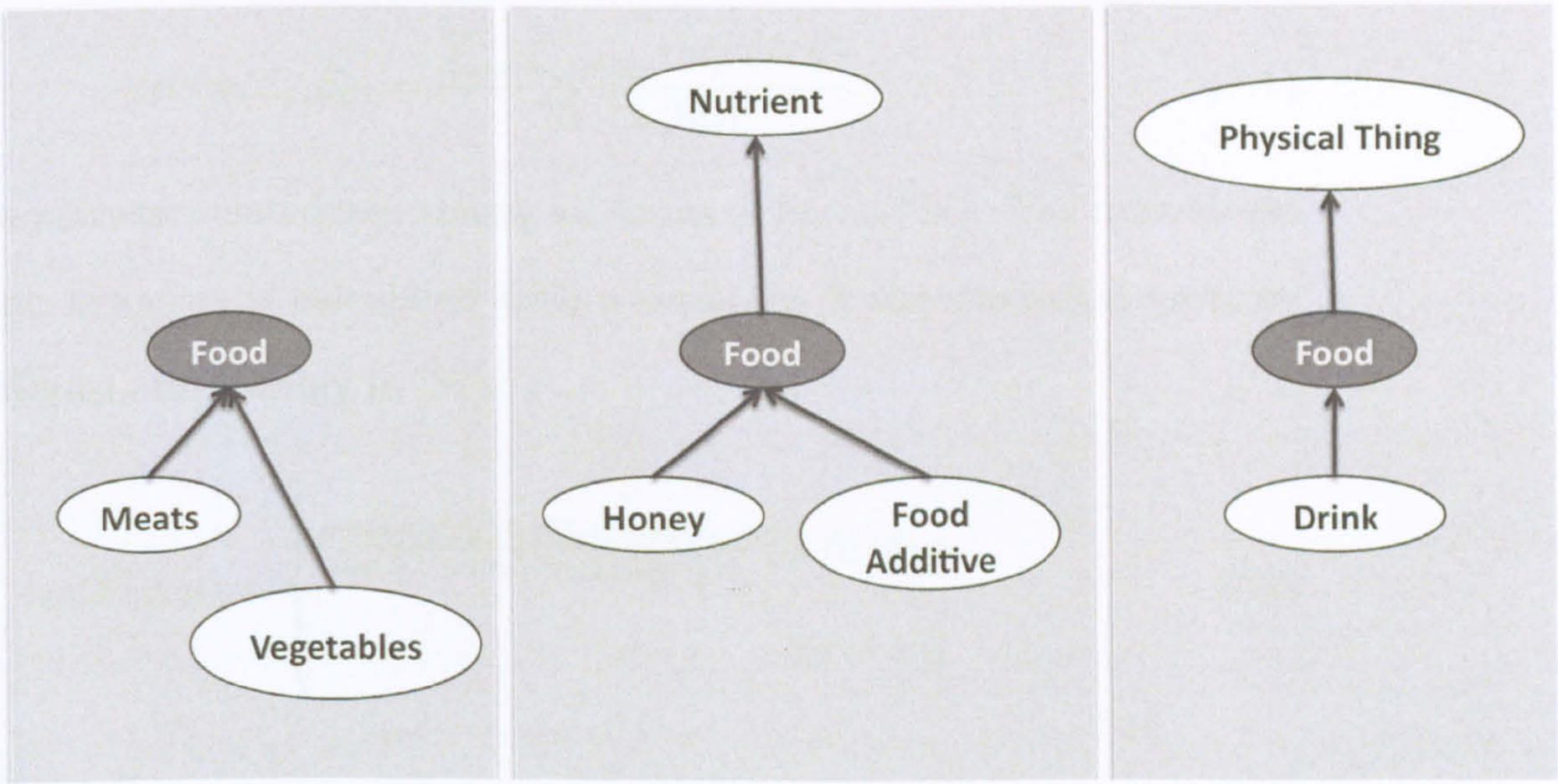


Figure 7.7: Three different senses for *Food*

the similarity measure did not cater for heterogeneous senses. For instance, not all the senses have supersenses or subsenses and only the lexical comparison of their neighbourhood may not be sufficient for the calculation of a similarity degree. In Figure 7.7 we present three senses created from semantic entities from online ontologies describing the concept of food. None of the three entities have further lexical information, e.g., synonyms, apart form their local name, which is `food`. In this case the lexical and structural information exploited in measure M4.2 are not sufficient to achieve adequate similarity value despite the fact that these senses refer to the same concept and should be merged into one.

We introduce the concept of **lexical neighbourhood** for the senses and use it in the devised similarity measure. The lexical neighbourhood of a sense is the vector of lexical information of its subsenses and supersenses. For example the lexical neighbourhoods of the senses in Figure 7.7 are {`meats`, `vegetables`}, {`nutrient`, `honey`, `food additive`} and {`physical thing`, `drink`}. We use the relatedness of the lexical neighbourhoods of the senses (vectors of terms) as an indicator of the senses' distance.

The relatedness of the lexical neighbourhoods ln_1 and ln_2 of the senses $S_1$and S_2 is

calculated as:

$$relLn(S_1, S_2) = \frac{\sum_{x \in \ln_1, y \in \ln_2} relT(x, y)}{|\ln_1| * |\ln_2|} \tag{7.1}$$

which is the mean pairwise relatedness among all terms of $\ln_1$ and $\ln_2$. The relatedness between two terms, x and y, is calculated with a variation of the statistical measure introduced by Cilibrasi and Vitaniy in [37]:

$$relT(x, y) = \begin{cases} \frac{max\{log(f_x), log(f_y)\} - log(f_{xy})}{log(N) - min\{log(f_x), log(f_y)\}} & if x \neq y \\ 1 & if x = y \end{cases} \tag{7.2}$$

In the measure of Cilibrasi and Vitaniy f_x and f_y represent the number of web documents when x and y occur individually and f_{xy} the number of documents where they occur together. We substitute the number of web documents with the number of folksonomy clusters where x and y occur. As discussed in Section 3.2, the tags belonging to the same cluster are related, so if f_{xy} is high (i.e., the number of clusters where x and y occur together), there is high probability that x and y are related. N is the total number of clusters associated with all tags of the tagspace.

The relatedness of lexical neighbourhoods, $relLn(s_1, s_2)$, is included in the measure M4.2 and the new sense similarity is calculated via the modified measure:

$$Sim(S_1, S_2) = W_L * SimL(S_1, S_2) + W_G * SimG(S_1, S_2) + W_{Ln} * RelLn(S_1, S_2) \tag{7.3}$$

where $simG(S_1, S_2)$ and $simL(S_1, S_2)$ are the graph similarity and lexical similarity. Note that the lexical similarity refers to string comparison of the sense's synonyms (in M4.2, this was the lexical similarity of labels and local names, which constitute the synonyms for senses). While the lexical similarity is calculated in the same manner as in M4.2, we modified the **graph similarity** as follows. Instead of lexically comparing

the neighbourhood of the two senses using string metrics, we used a semantic similarity measure (Wu and Palmer [125]) to cater for neighbour entities which are not identical, but they may synonymous or hierarchically related. For example, the appearance of *banana* as a subclass of *fruit*, and its appearance as a subclass of *food* would yield low graph similarity in M4.2 because the lexical similarity of *fruit* and *food* is low. Introducing semantic similarity for the comparison of parents (*fruit* and *food*) addresses this issue. This is because *food* subsumes *fruit*.

A **weighting function** is responsible for the adjustment of the weights w_G, w_L and w_{Ln} depending on the nature of S_1 and S_2. For instance, if S_1 has only supersenses and S_2 has only subsenses, then weights w_L and w_{Ln} are increased to compensate for the null value returned by $simG(S_1, S_2)$. This is a case of the senses of Figure 7.7. Also, if $simG(S_1, S_2)$ and $relLn(S_1, S_2)$ are lower than specified thresholds, this is an indication that the senses are likely to be dissimilar. As a result, the respective weights are lowered in order to avoid similarity values that would yield incorrect sense merging.

Note that $simG(S_1, S_2)$ represents the similarity of the senses neighbourhood e.g., "food is similar to fruit" based on the Wu and Palmer similarity on the WordNet hierarchy. $relLn(S_1, S_2)$ represents the statistical relatedness of the neighbourhood e.g., honey may not be similar to drink (either because they are not connected in the WordNet hierarchy or because one of them does not exist in it) but "honey is related to drink" because they frequently co-occur in clusters of related tags.

Sense Integration

If the similarity value of two senses S_1 and S_2 is higher than a threshold (decided using empirical experimentation) then a new sense is created containing the information of the two, $S_1 \bigcup S_2$. Consider the first two senses of food in Figure 7.7:

Sense: ***food*** $\xrightarrow{superSenseOf}$ ***http://..O1#meats***
O1 $\xrightarrow{superSenseOf}$ ***http://..O1#vegetables***
$\xrightarrow{isFoundIn}$ ***http://..O1#food***
" "

Sense: ***food*** $\xrightarrow{subSenseOf}$ ***http://..O2#nutrient***
O2 $\xrightarrow{superSenseOf}$ ***http://..O2#honey***
$\xrightarrow{superSenseOf}$ ***http://..O2#foodAdditive***
$\xrightarrow{isFoundIn}$ ***http://..O2#food***
" "

The product of their integration is:

Sense: ***food*** $\xrightarrow{subSenseOf}$ ***http://..O2#nutrient***
O1 $\xrightarrow{superSenseOf}$ ***http://..O1#meats***
O2 $\xrightarrow{superSenseOf}$ ***http://..O1#vegetables***
$\xrightarrow{superSenseOf}$ ***http://..O2#honey***
$\xrightarrow{superSenseOf}$ ***http://..O2#foodAdditive***
$\xrightarrow{isFoundIn}$ ***http://..O1#food***
$\xrightarrow{isFoundIn}$ ***http://..O2#food***
" "

We note that the level of abstraction for the subsenses of the new sense is variable, which can cause relation redundancy in the structure. For example, consider that a new sense is added, *sweetener* as a subsense of *food* and supersense of *honey*. Hence two relations exist between *food* and *honey*. One is explicit and another one is implicit, via *sweetener*. At this stage there is not enough information about which additional senses could be merged with the sense of *food*, and what other relations they could contribute to it. As a result, we do not further process the senses at this phase.

Figure 7.8 presents the physical output of the sense creation phase for the sense *Mosque*. This sense originates from three ontological entities and one WordNet synset. These are

```
@prefix flor: <http://flor.kmi.open.ac.uk/FLOR#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

flor:Sense_100527190844258_mosque
        flor:hasGloss
                "a Muslim place of worship",
                "A religious building where Islamic services and activities are
                held. Some of these are deinitely ModemShelterConstructions,
                but some are not.",
                "Islam a Muslim place of worship";
        flor:hasPart
                flor:_WN_3621419_noun,
                flor:_WN_3626641_noun;
        flor:hasProvenance
                "SW_WN_SW_SW";
flor:hasSynonym
                "masjid",
                "mosque",
                "musjid";
flor:isFoundIn
                flor:_WN_3646282_noun,
                <http://ontosem.org/#mosque>,
                <http://paoli.open.ac.uk/watson-cache#Mosque>;
flor:subSenseOf
                <http://ontosem.org/#religious-building>,
                <http://paoli.open.ac.uk/watson-cache #ReligiousBuilding>,
                <http://paoli.open.ac.uk/watson-cache#ReligiousStructure>;
a flor:Sense.
```

Figure 7.8: An example of RDF-encoded output for the sense of *Mosque*

related to the sense via the relation *flor:isFoundIn*. The value of the datatype relation *flor:hasProvenance* also demonstrates that the sense was created by three ontological entities (SW, "Semantic Web") and one WordNet entity. In addition it has two glosses and three synonyms which are contributed from the four semantic entities of origin. Finally we note that it relates to three entities, which are its supersenses, and two entities with the relation *flor:hasPart*.

Sense integration is the final process of the sense addition (Figure 7.1: step 6) and sense discovery phase. It returns the candidate senses for the tags which will be used for disambiguation in the next phase. The sense addition step performed in the scope of the semantic aggregation (Figure 7.1: step 13) also integrates senses to the repository using the similarity measure M7.3. In this case the output of sense addition is not a set of candidate senses for a tag but a sense used for the purposes of structure integration. We detail the semantic aggregation phase in Section 7.6.

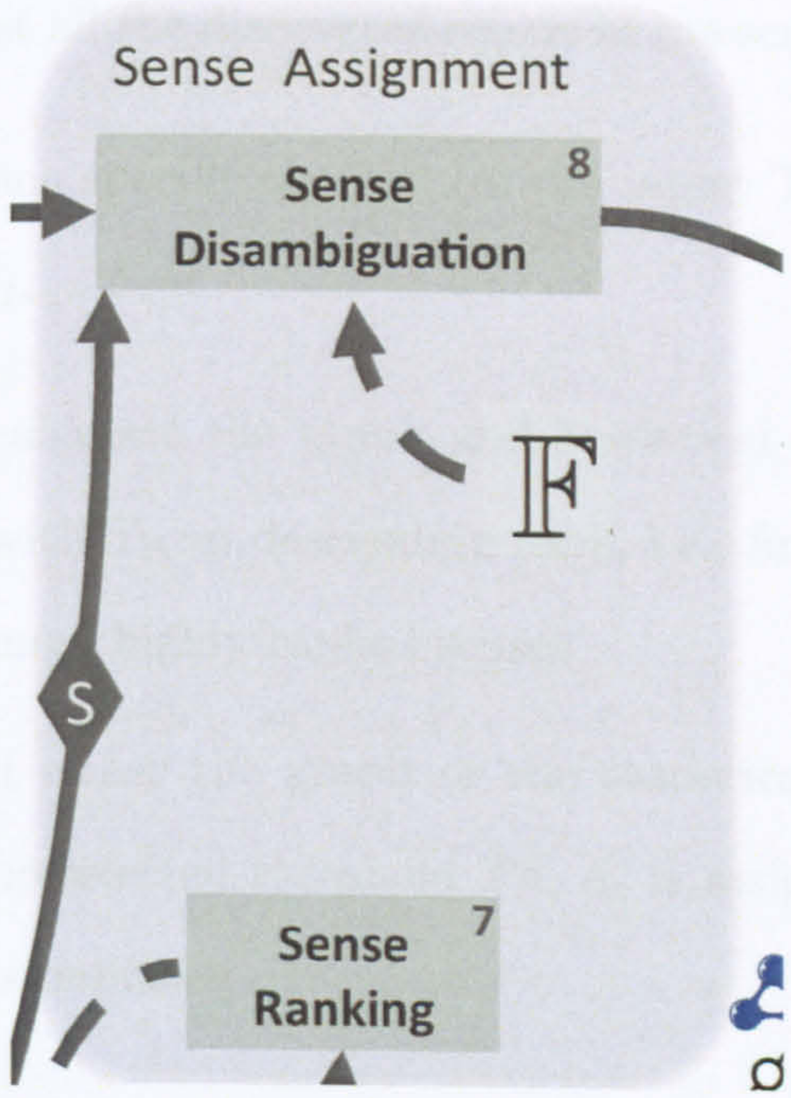


Figure 7.9: The Sense Assignment Phase

7.5 Sense Assignment

Sense discovery yields a set of candidate senses for each specific tag in the tagspace. The sense assignment phase is responsible for selecting the most appropriate sense for a specific tag taking into account the tagset it belongs to. In addition, it selects the sense, which is more valuable to the enrichment process. For example, an integrated sense originating from different ontologies and WordNet is usually richer than a sense that originates from a single ontology. This is because, integrated senses obtained using the similarity measure (described in the previous section) contain complementary information from different Knowledge Sources (see the example of the integrated sense of food). Also, a sense which originates from a Knowledge Source, from which more senses are extracted, is more likely to produce relations among the senses of the tagspace. Therefore we introduce a method which measures the value of candidate senses prior to disambiguation.

Tag disambiguation is performed as follows:

1. At first all the discovered senses in the sense repository are ranked (Section 7.5.1)

2. For each specific tag `R_t` (in the tagset T_R of resource R) with candidate senses S=$\{S_1,..., S_n\}$:

   (a) calculate the graph and statistical overlap (Section 7.5.2) of each $S_a \in$ S with T_R in descending rank, i.e., first calculate the overlap of T_R with the most highly ranked senses.

   (b) if either the graph or the statistical overlap of sense S_a is higher than a preselected threshold Th, S_a is assigned to `t` and the disambiguation for `t` terminates.

3. The disambiguation for `t` fails if neither the graph nor the statistical overlap measures exceed Th.

In the following we describe the functions of sense ranking and sense disambiguation in more detail.

7.5.1 Sense Ranking

The sense disambiguation step takes place after all candidate senses for the tagspace have been discovered. This fact allows for an overview of the candidate senses with respect to the following considerations:

- For each sense we are aware of its integration ratio, IR(S). This is the number of senses that have been integrated into one sense. For example, the integration ratio of the sense of *Mosque*, shown in Figure 7.8, is 4. A sense with higher integration ratio:

  1. originates from different Knowledge Sources. Therefore, it is possibly richer in lexical information and relations. The existence of variable relations from

heterogeneous sources is more likely to provide higher connectivity among the senses that belong to the output of FLOR-2.

2. is popular across different Knowledge Sources. The popularity of the sense reflects the fact that it is frequently used in (possibly) different domains and as a commonly used sense is likely to overlap well with the tagspace, which contains commonly used tags.

Therefore, senses with higher integration ratio IR(S) should be preferred when deciding which is the correct sense for a tag.

- For each sense we are aware of the similarity of senses used to create it, $Sim(S)$. The hypothesis is that a sense obtained by merging two highly similar senses is more likely to be correct than one that was obtained by merging two less similar senses. Therefore we give preference to the sense with the higher $Sim(S)$.

- We are aware of the popularity of each Knowledge Source, which is expressed by the number of senses that originate from it, i.e., how many entities of this Knowledge Source have been used for the creation of senses. The more popular the Knowledge Source, the more likely it is to provide better connectivity of senses, because the entities used to create them are more likely to be connected in the Knowledge Source of origin. As a result, a sense that originates from a popular Knowledge Source should be preferred. This is quantified with the expression $|\bigcup_{KS \in prov(S)} \mathcal{S}_{\text{KS}}|$ which is the number of senses ($\mathcal{S}_{\text{KS}}$) that originate from the Knowledge Sources of provenance of S. For example, consider sense S which is the result of merging three senses, created with three semantic entities originating from two ontologies, O1 and O2, and WordNet. In that case the provenance of S is $prov(S) = \{\text{O1}, \text{O2}, \text{WN}\}$. If WordNet has contributed a total of five candidate senses[5] then $|\mathcal{S}_{\text{WN}}| = 5$. Equally, consider that $|\mathcal{S}_{\text{O1}}| = 3$ and $|\mathcal{S}_{\text{O2}}| = 2$. Then

[5]Our previous experience showed that WordNet covers a high percentage of the tagspace. This means that the number of senses originating from WordNet could be higher than the number of all senses originating from ontologies. Therefore, to normalise the popularity of WordNet compared to the other ontologies we assign as $|\mathcal{S}_{\text{WN}}|$ not the actual number of senses originating from WordNet but the maximum value of senses obtained by one ontology.

$$|\bigcup_{KS\in\{O1,O2,WN\}} \mathcal{S}_{\text{KS}}| \;=\; 5+3+2=10.$$

Taking into account the above considerations, we define the following measure to calculate the rank of sense S.

$$rank(S) = IR(S) * Sim(S) * log\left(|\bigcup_{KS\in prov(S)} \mathcal{S}_{\text{KS}}|\right) * VBgr(S) \qquad (7.4)$$

VBgr(S) is used to increase the rank of S when this is created from variable backgrounds, i.e., a combination of ontologies and WordNet. The experiments in Chapter 6 showed that the knowledge in ontologies and WordNet is complementary therefore when integrated, it can provide a better connectivity of senses.

7.5.2 Sense Disambiguation

This is the last step (Figure 7.1: step 8) of sense assignment, it disambiguates the candidate senses of a tag `t` and selects the sense which is more appropriate for the context of `t`. This process is designed to satisfy Rq_1, which is to exploit semantic and statistical information for the purposes of sense disambiguation. We calculate the graph-based and the statistical overlap of the candidate senses with the tagset of `t` and select the highest ranked sense that exceeds a predefined threshold.

Graph-based Overlap

Graph based overlap allows for relation based disambiguation and exploits the relations among the candidate senses of a tagset T. Our hypothesis is that among the candidate senses of `t`, the one that is better connected via formal relations with the senses of the other tags in T is the most likely to represent the correct meaning for `t`. We represent this degree of connectedness as follows. For each candidate sense S of `t`, which belongs

to tagset T, we calculate its graph-based overlap with T using the following measures:

$$O_{\mathrm{G}}(S,T) = \sum_{S_{\mathrm{i}} \in T} \sum_{e \in S,\ e_{\mathrm{i}} \in S_{\mathrm{i}}} d_{\mathrm{G}}(e, e_{\mathrm{i}}) \tag{7.5}$$

$$d_{\mathrm{G}}(e, e_{\mathrm{i}}) = \begin{cases} \frac{1}{|p(e,e_{\mathrm{i}})|} & \text{if there is a path that connects e and } e_{\mathrm{i}} \\ 0 & \text{if no such path exists OR} \\ 0 & \text{e and } e_{\mathrm{i}} \text{ do not belong to the same Knowledge Source.} \end{cases} \tag{7.6}$$

S_i represents the other candidate senses of the tags of T, while e and e_i are semantic entities which were used to create the senses S and S_i. Using these original entities we calculate $d_{\mathrm{G}}(e, e_{\mathrm{i}})$ which represents their distance in the graph of the Knowledge Source of origin. This measure also represents the semantic distance between the senses S and S_i with $|p(e, e_{\mathrm{i}})|$ being the length of the connecting path between e and e_i in the Knowledge Source. The connecting path may include all types of relations and is not restricted to subsumption. In the cases when e and e_i neither belong to the same Knowledge Source nor are they connected to each other, $dG(e, e_{\mathrm{i}}) = 0$.

This formula caters for all senses regardless of their provenance and whether they have been merged beforehand. The distance $d_{\mathrm{G}}(e, e_{\mathrm{i}})$ is calculated for all pairwise combinations of the entities of S and S_{i}. If the senses have been merged and some of their entities originate from the same Knowledge Source then the value of $d_{\mathrm{G}}(e, e_{\mathrm{i}})$ would be non null, therefore contributing to the distance of the two senses.

The graph-based overlap of the candidate sense S with T, $O_{\mathrm{G}}(S,T)$, is the sum of $d_{\mathrm{G}}(e, e_{\mathrm{i}})$ among all entities e of S and all entities e_i of the other candidate senses of the tags in T, S_i. This type of disambiguation takes into account all the candidate senses assigned to the tags of T but does not consider the rest of the tags for which no

candidate sense is found.

Statistical Overlap

In contrast to graph-based overlap, the statistical overlap of a candidate sense S with the tagset T of `t` takes into account all tags of T regardless if they have been assigned candidate senses S_i or not. The statistical disambiguation does not exploit the relations of the candidate sense S with S_i. It exploits its lexical neighbourhood, $ln(S)$, by calculating the relatedness of the neighbourhood with the tags of the tagset T. The relatedness of $ln(S)$ with the tags of the tagset is called *statistical overlap of* S *with* T and is calculated using the following measure:

$$O_{\mathrm{S}}(S,T) = \frac{\sum_{x\in\ \mathrm{ln(S)}, y\in\ \mathrm{T}} relT(x,y)}{|\mathrm{ln(S)}| * |\mathrm{T}|} \tag{7.7}$$

$relT(x,y)$ is the relatedness between y and x, as calculated by measure M7.2.

This statistical disambiguation caters for cases where the relation based disambiguation fails due to the heterogeneous origin of the senses in a tagset (which does not allow for connecting paths in one Knowledge Source). As in relation based disambiguation, the most strongly connected sense to all the tags of the tagset is selected to define the meaning of the tag.

If sense assignment successfully disambiguates tag `R_t` (the specific occurrence of `t` in the tagset of resource R) and assigns to it sense S, the following relation (represented by the couple {`t`, S} in Figure 7.1)

$$\texttt{R\_t} \xrightarrow{hasDefinition} S.$$

is appended to the semantic structure, which is the output of FLOR-2. This process, sense disambiguation, and as a result sense assignment, fails when both $O_{\mathrm{G}}(S,T)$ and

$O_S(S,T)$ fall under the specified thresholds.

Figure 7.10 shows the output of sense assignment to the tags of the tagset 24768. `24768_exhaustion` and `24768_sitting` are the only tags not assigned a sense. This is because the overlap of their candidate senses with the other tags of the tagset is lower than the preselected threshold. Hence graph overlap is zero because there are no formal relations connecting the senses of *exhaustion* and *sitting* with the senses of the other tags in 24768. In addition, the statistical overlap of the two tags with the rest of the tagset is low. Indeed, while {`mosque`, `tourist`, `istanbul`, `turkey`} provide a good disambiguation context for each other, they are not sufficiently related to {`exhaustion`, `sitting`} in order to facilitate their disambiguation.

7.6 Semantic Aggregation

This is the last phase of FLOR-2 and is responsible for the creation of the sense structure $\mathcal{S}$, which represents the meanings of tags in $\mathcal{T}$ and their relations. The previous phase, sense assignment, populates the output with relations among tags and senses. However, the relations between the senses, and thus between the tags, are not specified yet. In Section 7.4 we described how the newly created senses relate to the neighbour entities of their original semantic entity. In this phase, the relations among senses and semantic entities are leveraged to relations between senses. This is achieved during relation definition (Figure 7.1: step 9) where the existing senses are connected with each other (Section 7.6.1). Yet, to produce a connected graph we introduce superordinate senses that do not exist in the tagspace[6] but are used to summarise the existing senses (Figure 7.1: steps 10 to 14, Section 7.6.2). This is important in order to create a connected hierarchy. For example, if the senses *Italy*, *Hungary* and *Slovenia* already exist in $\mathcal{S}$, the new sense *Country*, is introduced as well as its relations to the existing

[6]If there were tags in the tagspace with the meaning of such superordinate senses the later would have been discovered and created by steps 3 to 6

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#>.
@prefix flor: <http://flor.kmi.open.ac.uk/FLOR#>.
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>.

flor:Tag_24768_exhaustion a flor:Tag;
        rdfs:label
                "exhaustion".
flor:Tag_24768_istanbul a flor:Tag;
        flor:hasDeflnition
                flor:Sense_100527061916718;
        rdfs:label
                "istanbul".
flor:Tag_24768_mosque a flor:Tag;
        flor:hasDeflnition
                flor:Sense_100527190844258_mosque;
        rdfs:label
                "mosque".
flor:Tag_24768_sitting a flor:Tag;
        rdfs:label
                "sitting".
flor:Tag_24768_tourists a flor:Tag;
        flor:hasDeflnition
                flor:Sense_100527191014928_tourist;
        rdfs:label "
                tourists".
flor:Tag_24768_turkey a flor:Tag;
        flor:hasDeflnition
                flor:Sense_10052744512923;
        rdfs:label
                "turkey".
flor:TaggedResource_24768
        flor:isTaggedWith
        flor:Tag_24768_exhaustion,
        flor:Tag_24768_istanbul,
        flor:Tag_24768_mosque,
        flor:Tag_24768_sitting,
        flor:Tag_24768_tourists,
        flor:Tag_24768_turkey;
        a flor:TaggedResource.
```

Figure 7.10: An example of RDF-encoded output for the enriched tagset "24768"

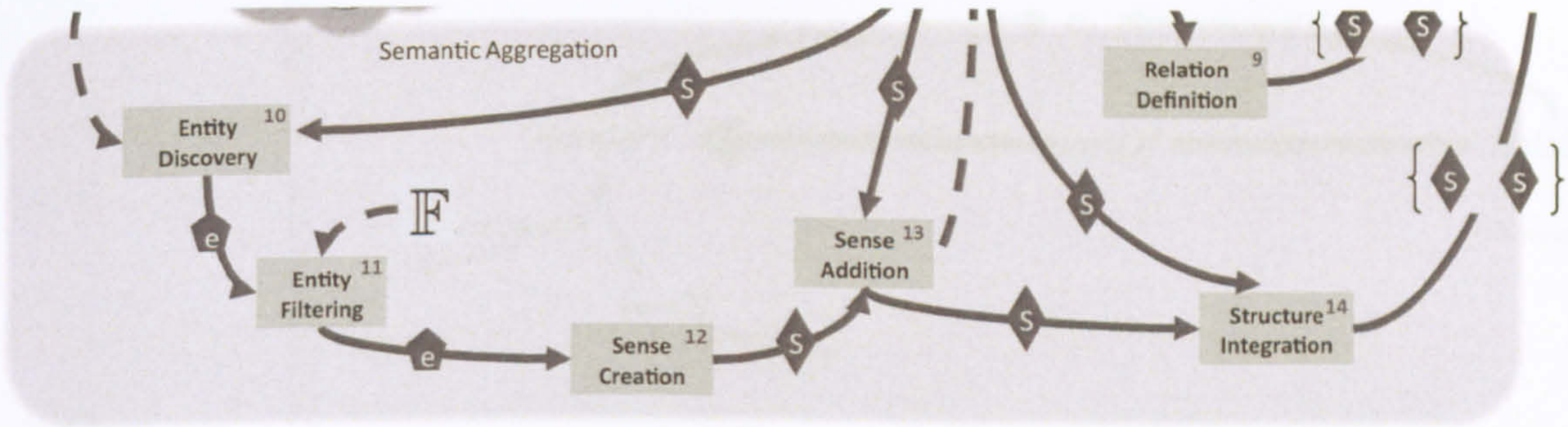


Figure 7.11: The Semantic Aggregation Phase

three senses that represent instantiations of countries.

7.6.1 Sense Relation Definition

The previous processes of FLOR-2 (steps 1 to 8) return two types of relations between semantic entities and senses. The first type represents the fact that the sense S was created by entity e and is:

$$S \xrightarrow{isFoundIn} e$$

The second type of relations refers to the ones described in Section 7.4.3, which include subordinate, superordinate and meronymy relations. These relations emerged from the relations of the entity e, which was used for the creation of a sense S, with its neighbour semantic entities, e'. At that stage of sense creation, the relations among e and e' were translated to relations among S and e':

$$S \xrightarrow{subSenseOf} e', S \xrightarrow{instanceOf} e', \ldots$$

The above two types of relations are used for the integration of senses that have been assigned[7] to tags in the disambiguation phase. The process of relation definition among assigned senses is:

[7] Although more candidate senses have been discovered, if these were not used to define the meaning of a tag in the tagspace, they are not considered by the semantic aggregation phase.

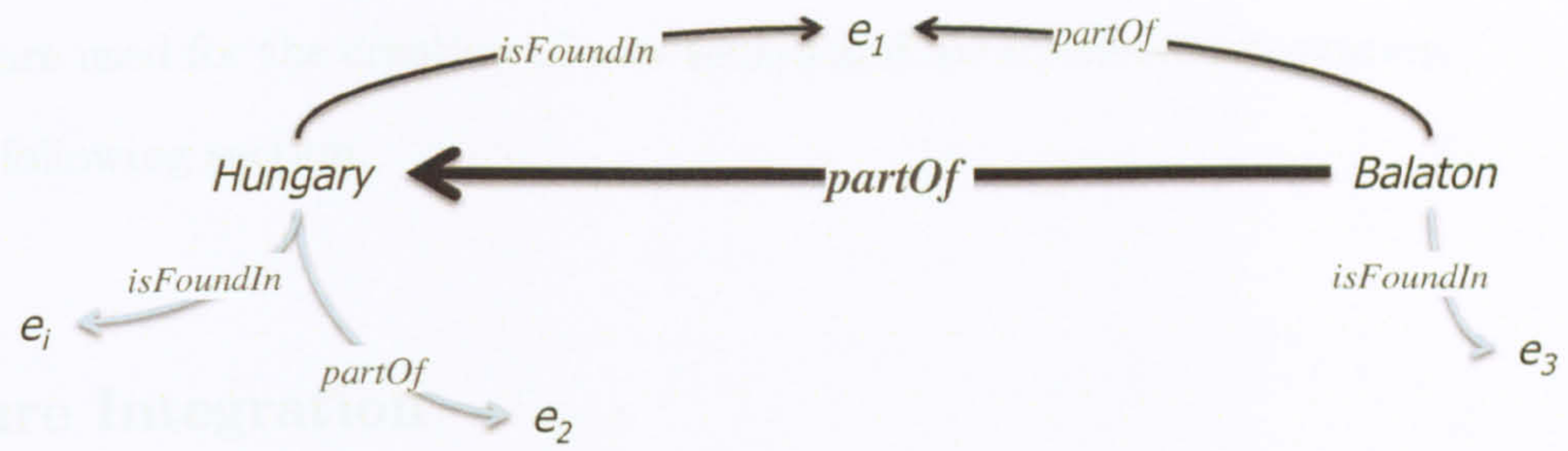


Figure 7.12: Relation Definition Process

1. For each sense S, extract the semantic entities, e', with which S is related via a specific relation. For example, all e' which relate with S via the relation *subSenseOf*: $S \overset{subSenseOf}{\longrightarrow} e$'.

2. For all the semantic entities, e', locate all senses S' that have been created from e', i.e., $S' \overset{isFoundIn}{\longrightarrow} e$'.

3. Connect S and S' with the relation of S and e'. For example, $S \overset{subSenseOf}{\longrightarrow} S'$

4. Repeat actions 1-3 for all the different types of relations between S and e' (*perSenseOf*, *partOf*, *instanceOf*, and so on).

The process is exemplified in Figure 7.12 where two senses share different types of relations with semantic entities. For example, the relations:

$$Hungary \overset{isFoundIn}{\longrightarrow} e_1 \text{ and } e_1 \overset{partOf}{\longleftarrow} Balaton$$

produce:

$$Hungary \overset{partOf}{\longleftarrow} Balaton$$

The process is repeated for all assigned senses and all the entities e' with which they relate. In the case when some of e', have not been already used to create another sense (i.e., there is no sense S' for which $S' \overset{isFoundIn}{\longrightarrow} e$'), this means that the concept defined by e' does not appear in the input tagspace $\mathcal{T}$. Yet, the entities e' that are subsumed by

the existing senses, are used for the creation of new senses and for structure integration as described in the following section.

7.6.2 Structure Integration

This is the last step of the algorithm which performs the integration process over the senses discovered for a tagspace. For example, consider that the senses *Italy*, *Austria* and *Hungary* have been assigned to tags in the tagspace. If the input tagspace $\mathcal{T}$, does not contain the tag `country`[8], then the process of sense discovery (Figure 7.1: steps 3 to 6) has not been triggered for this tag, and there is no sense such as *Country* in the repository. However, importing this new sense and its relations to the existing senses *Austria*, *Hungary* and *Italy*, would render the output $\mathcal{S}$ a connected structure.

The process of structure integration takes as input all the senses S_n that are not connected to another sense with the hierarchical relations *subSenseOf* and *instanceOf*, and imports their superordinate senses. For all senses that have no superordinate senses, S_n:

1. extract entities e_p which are related to S_n via the relations *subSenseOf* or *instanceOf* (Figure 7.1: step 10).

2. For each e_p:

   (a) create sense S_p according to the process described in Section 7.4

   i. Entity filtering (Figure 7.1: step 11), which validates that e_p is meaningful to folksonomy enrichment.

   ii. Sense creation (Figure 7.1: step 12), creates S_p.

   iii. Sense addition (Figure 7.1: step 13), which ensures that S_p is a new sense and there is no other sense with similar meaning to it in the repository.

[8]Or any other tag which is a synonym of `country`

In the case when a sufficiently similar sense to S_p exists, e.g., S_s, these two are integrated (in the same manner described in Section 7.4.4) and S_p becomes the integrated sense $S_p = S_p \bigcup S_s$.

(b) Add the relation $S_n \overset{subSenseOf/instanceOf}{\longrightarrow} S_p$ to the output $\mathcal{S}$.

3. Repeat actions 1-2 for S_p until the subsumption path of entities has reached the root of the Knowledge Source.

4. The process terminates when all S_n that are not connected to superordinate senses are integrated in the hierarchy.

For example, consider the aforementioned three senses which are not connected to other senses with subsumption relations, *Italy*, *Hungary* and *Austria*. Structure integration is performed as follows. The superordinate semantic entities of $S_n = Italy$ are extracted, for example, the entity $e_p = http://Ontology1.com\#Country$ is related to *Italy* via the relation *instanceOf*. e_p is used for the creation of $Country_p$ which is a new sense (no other sense in the repository is sufficiently similar to it). The relation:

$$Italy \overset{instanceOf}{\longrightarrow} Country_p$$

is added to the structure and the same process is repeated for $Country_p$, which yields a new superordinate sense, $Region_p$, and relation:

$$Country_p \overset{subSenseOf}{\longrightarrow} Region_p$$

The process continues until no more semantic entities are discovered in the path of e_p in its Knowledge Source of origin (Ontology1), and adds the following path in $\mathcal{S}$:

$$Italy \overset{instanceOf}{\longrightarrow} Country_p \overset{subSenseOf}{\longrightarrow} Region_p \overset{subSenseOf}{\longrightarrow} Location_p$$

Then the process is repeated for *Austria* and *Hungary*. The semantic entities that are subsumed by *Hungary* are extracted, for example the entity $e_p = http://Ontology2.com\#Count$

Then $e_{\acute{p}}$ is used for the creation of $Country_{\acute{p}}$. However, during sense addition (Figure 7.1: step 13), $Country_{\acute{p}}$ and $Country_p$ demonstrate high relatedness, thus instead of adding a new sense, $Country_{\acute{p}}$, we integrate it with the existing $Country_p$, $Country_p = Country_p \bigcup Country_{\acute{p}}$ and add the subsumption relation:

$$Hungary \overset{instanceOf}{\longrightarrow} Country_p$$

The steps 10-13 (of Figure 7.1) of the semantic integration are then omitted for $Country_{\acute{p}}$ as it has already been integrated into the existing structure through its merging with $Country_p$.

7.7 Summary

In this chapter we described the overall procedure and individual processes of FLOR-2 as presented in Figure 7.1. We showed how, given an input tagspace $\mathcal{T}$, we obtained a structure of senses $\mathcal{S}$, which define the meaning of tags in $\mathcal{T}$, and the relations between these senses. In the next chapter we describe the evaluation of the algorithm in terms of sense assignment correctness, i.e., what percentage of assigned senses is correct, tagspace coverage, i.e., what percentage of tags is assigned a correct sense, and connectivity of obtained structure.

Chapter 8

Evaluating the Enrichment of Tagspaces

In this chapter we evaluate the FLOR-2 enrichment algorithm in terms of sense assignment correctness, connectivity of sense spaces and tagspace coverage. We use two datasets one of which is the dataset used to evaluate FLOR-1. With this we perform a comparative evaluation of the two versions of FLOR in terms of tagspace coverage and sense assignment correctness.

8.1 Introduction

In this chapter we evaluate our enrichment algorithm from three different perspectives:

- **Evaluation of the correctness of sense assignment**, i.e., decide whether the senses that FLOR-2 assigned to tags are correct given their context (Section 8.2).

- **Evaluation of the degree of connectivity** of the structure generated by the semantic aggregation phase, i.e., decide if the relations discovered between the senses constitute a semantic layer with a high degree of connectivity (Section 8.3).

- **Evaluation of tagspace coverage**, i.e, measure the percentage of correctly enriched tags compared to the total tags of the input tagspace and identify the reasons for non-coverage of the rest of the tags (Section 8.4).

In the following section we give an overview of the experiments carried out and present the datasets used for the purposes of this evaluation.

8.1.1 Experimental Setup

In order to evaluate the enrichment algorithm with respect to the three tasks introduced above, we use two different datasets. Table 8.1, summarises their characteristics. Dataset A has already been used to evaluate FLOR-1 (Section 4.6). The reason for selecting the same dataset is to compare the relative improvement of FLOR-2 with respect to sense assignment correctness and tagspace coverage (note that FLOR-1 did not support semantic aggregation). Dataset A comprises 250 resources randomly selected from Flickr with 2819 generic tags and 4242 specific tags[1] (the distinction between generic and specific tags was given in Definition 2 from Section 3.2). Dataset B is to the best of our knowledge the only established folksonomy dataset used for evaluation and was used in the experiments of Chapter 6. It is the MIRFLICKR-25000 collection [61] and was introduced for the purposes of image retrieval evaluation [10]. For this reason the images in this collection are of high quality and were selected based on their interestingness[2].

We have enriched the two datasets using FLOR-2 and used human evaluations to asses the correctness of the sense assignment. On the basis of this result we then evaluate the semantic aggregation and tagspace coverage. Due to the large scale of the FLOR-2 output, we use the following convention throughout the evaluation. The evaluation

[1]The generic tags occur only once in the tagspace, e.g., `apple`, while the specific occur as many times as the number of resources they tag. For example, in Dataset A there can be potentially 250 instances of `apple` as a specific tag if all resources are tagged with `apple` (see Definition 2).

[2]`http://www.flickr.com/explore/interesting`

	Dataset A	Dataset B
Number of Resources	250	25000
Number of Generic tags	2819	69099
Number of Specific Tags	4242	223537

Table 8.1: Evaluation datasets

itself has been carried out on a randomly selected sample of the set of sense assignments. Without loss of generality we assume that the evaluation results from the sample are a meaningful approximation of the result for the overall population.

In the following sections we describe the evaluation strategies for each of the three tasks.

8.2 Evaluation of Sense Assignment

The goal of the sense assignment evaluation is to decide if the sense S assigned to tag `t` is appropriate and conveys the meaning of `t` in the context of the tagset of resource R. This is the output produced by the process of sense disambiguation (Figure 7.1: step 8) and is a sense assignment sa represented by the relation:

$$\texttt{R\_t} \xrightarrow{hasDefinition} S$$

To evaluate the correctness of this assignment we devised the method presented in [132] and used the help of a group of human judges (N). From the set of all sense assignments produced by FLOR-2, A, we randomly selected a subset $SA : A \supset SA$. We then asked each judge to asses if each sense assignment $sa \in SA$ was correct or not.

Each judge is given a set of sense assignments in the manner presented in Table 8.2. The first column contains the tag whose sense assignment is under evaluation. The

	Tag	Tagset	Sense
sa_1	light	window, blue, *light*, orange, warmth	GrowthCondition, status, condition, illumination, lighting
sa_2	fire	blue, demon, pipaugust, hand, *fire*, lowlight, devil	Phenomena, firestorm, natural-event
sa_3	alberta	winter, *alberta*, path, calgary	canadian_province

Table 8.2: Evaluation input example for sense assignment.

second column contains the tag in its original context, i.e., among the other tags in its tagset. The last column contains a set of terms describing the assigned sense. These terms are the semantic neighbourhood of the sense, i.e., synonyms, supersenses and subsenses.

The evaluators were asked to asses whether the meaning conveyed from the set of terms in the last column was the correct meaning for the tag given its context tagset. They could answer either "yes", "no" or "unsure" if the assignment is correct, incorrect or they can not make a judgement. As a result, for each sense assignment, sa, we obtain a tuple of responses, $R =< r_{J_1}, r_{J_2}, ..., r_{J_N} >=< yes, no, unsure, ..., unsure >$, one response per judge.

To calculate the global consensus on the correctness of sa, i.e., if the majority of judges believe it is correct or incorrect, we construct $\acute{R} =< \acute{r}_{J_1}, \acute{r}_{J_2}, ..., \acute{r}_{J_N} >$, as follows:

- If r_{J_i}= "yes" then $\acute{r}_{J_i} = 1$
- If r_{J_i}= "no" then $\acute{r}_{J_i} = 0$
- If r_{J_i}= "unsure" then $\acute{r}_{J_i} = 0.5$

For example, if the response tuple for sa is $R =< yes, no, unsure, yes, yes >$, we obtain: $\acute{R} =< 1, 0, 0.5, 1, 1 >$. We then use $\acute{R}$ to calculate the **global degree of correctness** for sa, $gdc(sa)$ as follows:

$$gdc(sa) = \sum_{\acute{r}_{Ji} \in \acute{R}} \acute{r}_{Ji}$$

For example, $gdc(sa) = 1 + 0 + 0.5 + 1 + 1 = 3.5$.

Once $gdc(sa)$ is known for all sense assignments, we then decide which assignments are globally correct, which are globally incorrect and for which there is no global consensus. The rules of Table 8.3 are applied for the categorisation of the sense assignments:

Condition	sa Judgement
$gdc(sa) \geq t_C$	"correct"
$gdc(sa) \leq t_I$	"incorrect"
$t_I < gdc(sa) < t_C$	"undecided"

Table 8.3: Conditions for judging the global consensus for sa

t_C and t_I are thresholds used to decide if $gdc(sa)$ reflects a global correctness or incorrectness of sa. These thresholds are decided based on the number of judges N (see Sections 8.2.1 and 8.2.2)

$$\begin{aligned} &SA \supset SA_C = \bigcup sa_i \quad \forall sa_i \in SA : sa_i \text{ "correct"} \\ &SA \supset SA_I = \bigcup sa_i \quad \forall sa_i \in SA : sa_i \text{ "incorrect"} \\ &SA \supset SA_U = \bigcup sa_i \quad \forall sa_i \in SA : sa_i \text{ "undecided"} \end{aligned}$$

Table 8.4: Correct (SA_C), incorrect (SA_I) and undecided (SA_U) sense assignments

Using the strategy described above and the rules of Table 8.4, we obtain the sets of correct, incorrect and undecided sense assignments in SA. We use SA_C and SA_I to decide the ratio, r_{SA} of correct sense assignments in SA as follows:

$$r_{SA} = \frac{|SA_C|}{|SA_C \bigcup SA_I|} \tag{8.1}$$

SA is a randomly selected subset of A. Thus, we can assume without loss of generality, that the ratio of correct sense assignments in A, r_A can be approximated by r_{SA}, i.e., $r_A \simeq r_{SA}$. Yet, the number of correct sense assignments is equal to the number of

specific tags correctly enriched. As a result, the ratio presented in M8.1 represents the **precision of FLOR-2** in terms of tag enrichment.

In the following sections we detail the evaluation of the sense assignment of FLOR-2 on two datasets A and B.

8.2.1 Experiment A

To evaluate the correctness of sense assignment to the tags of Dataset A we randomly selected SA with $|SA| = 300$ and asked a group of $N = 4$ volunteers (postgraduate and postdoctoral researchers) to judge the correctness of the assignments. Table 8.5 contains the numbers of individual responses for each of the four judges. J_{A1} judged 262 sense assignments as correct, 29 as incorrect and she could not make a judgement for 9 of them.

	J_{A1}	J_{A2}	J_{A3}	J_{A4}
Yes	262	249	230	210
No	29	33	54	33
Unsure	9	18	18	57

Table 8.5: Experiment A: Individual responses of the four judges

For each sense assignment $sa \in SA$ we transformed the response tuple R to $\acute{R}$ and calculated the $gdc(sa)$. In order to categorise each sa we calculated the thresholds t_C and t_I using the rules of Table 8.6, which apply when N=4. We set $t_C = 3$ and $t_I = 1$. If there is at most one negative judge and everyone else is positive, we consider that there is enough evidence to support the global correctness of the assignment. Equally, if there is at most one positive judge we consider the assignment globally incorrect. This is because we were interested in obtaining a strong global consensus. As a result, all sa with $1 < gdc(sa) < 3$ are considered undecided.

sa is globally correct when:		
More than half of the judges are positive	$\acute{R} =< 1, 1, 1, 0 >$	$gdc(sa) = 3$
Half of the judges are positive and half are unsure (none is negative)	$\acute{R} =< 1, 1, 0.5, 0.5 >$	$gdc(sa) = 3$
sa is globally incorrect when:		
More than half of the judges are negative	$\acute{R} =< 1, 0, 0, 0 >$	$gdc(sa) = 1$
Half of the judges are negative and half are unsure (none is positive)	$\acute{R} =< 0, 0, 0.5, 0.5 >$	$gdc(sa) = 1$

Table 8.6: Rules for deciding the global correctness thresholds t_C and t_I for $N = 4$

Calculating the $gdc(sa)$ for all $sa \in SA$ we used the rules of Tables 8.3, 8.4 and 8.6 to obtain the subsets of SA as follows:

- $|SA_C| = 241$
- $|SA_I| = 17$
- $|SA_U| = 42$

Applying the above values measure to M8.1 we calculate the approximate correctness in sense assignment for Dataset A as:

$$r_A \simeq r_{SA} = \frac{|SA_C|}{|SA_C \bigcup SA_I|} = \frac{241}{258} = 0.934$$

Although this is an approximation of the overall precision for FLOR-2, it is a very close value to the one obtained from the experiments with FLOR-1, 0.93. This is a satisfactory result, given that FLOR-2 was not created to improve precision, but to improve coverage of the tagspace. Although we substituted the strict WordNet-based disambiguation methods with hybrid graph and statistical disambiguation, the precision rate remained the same. Finally, the minimum agreement for all evaluators was calculated as 0.71. This reflects the number of globally correct sense assignments with $gdc(sa) = 4$.

8.2.2 Experiment B

For the evaluation of FLOR-2 on Dataset B we also selected randomly a subset $|SA| = 300$ (see Appendix C for the complete set of sense assignments and the individual evaluations of the five judges). A different group of $N = 5$ volunteers, also postgraduate and postdoctoral researchers, were the judges in this experiment. We present the individual responses per judge in Table 8.7.

	J_{B1}	J_{B2}	J_{B3}	J_{B4}	J_{B5}
Yes	272	266	244	239	220
No	24	20	35	25	48
Unsure	4	14	21	36	32

Table 8.7: Experiment B: Individual responses of the five judges

For each sense assignment $sa \in SA$ we transformed the response tuple R to $\acute{R}$ and calculated the $gdc(sa)$. In order to categorise each sa we calculated the thresholds t_C and t_I using the rules of Table 8.8[3] and set $t_C = 3.5$ and $t_I = 1.5$. As a result, all sa with $1.5 < gdc(sa) < 3.5$ belong to the set of senses for which no global consensus has been achieved.

sa is globally correct when:		
More than half of the judges are positive and at least one is unsure	$\acute{R} =< 1, 1, 1, 0.5, 0 >$	$gdc(sa) = 3.5$
There are at least two positive and no negative judges	$\acute{R} =< 1, 1, 0.5, 0.5, 0.5 >$	$gdc(sa) = 3.5$
sa is globally incorrect when:		
More than half of the judges are negative and at least one is unsure	$\acute{R} =< 1, 0.5, 0, 0, 0 >$	$gdc(sa) = 1.5$
There are at least two negative and no positive judges	$\acute{R} =< 0, 0, 0.5, 0.5, 0.5 >$	$gdc(sa) = 1.5$

Table 8.8: Rules for deciding the global correctness thresholds t_C and t_I for $N = 5$

To calculate the $gdc(sa)$ for all $sa \in SA$ we used the rules of Tables 8.3 and 8.4 to

[3]These rules apply when N=5

obtain the subsets of SA as follows:

- $|SA_C| = 252$
- $|SA_I| = 17$
- $|SA_C| = 31$

Applying the above values to measure M8.1 we calculate the approximate correctness in sense assignment for Dataset B as:

$$r_A \simeq r_{SA} = \frac{|SA_C|}{|SA_C \bigcup SA_I|} = \frac{252}{269} = 0.936$$

Although Dataset B is different to A and the results were evaluated by a different group of judges, the value of sense assignment correctness is consistent with the values we obtained both for Dataset A and also in the experiments with FLOR-1 (Section 4.6). This is a satisfactory result given that, as mentioned before, the main goal of FLOR-2 was to improve tagspace coverage rather than precision. In the same line with the results of the experiment with Dataset A, the minimum agreement among the evaluators for senses with $gdc(sa) = 5$ was 0.72.

8.3 Evaluation of Semantic Aggregation

Here we follow an evaluation strategy based on the measures M3.2 to M3.5, defined in Section 3.6.1, which evaluate the structure in terms of subsenses, supersenses and synonyms. Table 8.9 shows the values obtained for the two datasets. Although measures M3.4 and M3.5 were defined to measure the number of subsenses and supersenses, in this occasion we include all subordinate and superordinate senses i.e., the senses connected with the relations *flor:hasInstance* and *flor:instanceOf*. In addition, for the

	Measure	Dataset A	Dataset B
M3.2	$\overline{\lvert syn(\mathcal{S}_{\mathrm{KS}})\rvert}$	2.3	2.2
M3.4	$\overline{\lvert sub(\mathcal{S}_{\mathrm{KS}})\rvert}$	1.5	1.8
M3.5	$\overline{\lvert sup(\mathcal{S}_{\mathrm{KS}})\rvert}$	3.0	2.9

Table 8.9: Quantitative results of the enrichment evaluation

evaluation of measures M3.2, M3.4 and M3.5 we used the senses which were correctly assigned to tags (using the results of the sense assignment evaluation).

If we compare the values of Table 8.9 with the ones of Table 6.1 (presents the results of sense richness in the structures created from WordNet and ontologies) we note that the senses present a quite similar number of synonyms (2.2-2.3). This is justified by the fact that the senses created in this evaluation originate from the same Knowledge Sources as the senses presented in Table 6.1.

In Table 6.1 we calculate the mean number of all subsenses and supersenses for each sense but in Table 8.9 we only calculate those which are assigned to tags. This justifies the difference in the number of subsenses for WordNet-derived senses (Table 6.1: 2.7), to the number of subsenses presented in Table 8.9 (1.5 and 1.8). Not all WordNet hyponyms of a synset were transformed by FLOR-2 to a sense which was then assigned to a tag of Dataset A. In addition, the number of subsenses for Dataset B is larger than the number of subsenses for Dataset A. This is because, given the magnitude of Dataset B, the probability of existence of tags in its tagspace which are connected to subordinate senses is higher.

The mean values for supersenses are larger in Table 8.9 in contrast to Table 6.1. This is due to the fact that in FLOR-2 supersenses are added to the structure in order to provide common ancestors for the existing senses and they are not required to link to a tag. Furthermore, the number of supersenses is larger because of the sense merging process which integrates senses with different parents into one sense (see example of food in Figure 7.7) and for each parent it adds its ancestors to the hierarchy. Finally,

measure M3.3, which represents the mean number of candidate senses for the tags of the tagspace, is lower for Dataset B because of the larger number of tags.

We evaluate the number of relations between the senses using the following strategy. Consider $\mathcal{E}$, which is the set of senses that were connected to each other during the relation definition step (Figure 7.1: step 9) and $\mathcal{A}$ the senses which were connected to superordinate senses during the structure integration step (Figure 7.1: step 14). As shown in Table 8.10, 46% of the correctly assigned senses to tags of Dataset A are connected to existing senses while for 11% of the senses there was no relation to the rest of the structure. For Dataset B 71% of the senses were connected during the relation definition step while only 8% were not connected to any other sense.

Number of Senses		A	B
connected during the relation definition step	$\mathcal{E}$	46%	71%
connected during the structure integration step	$\mathcal{A}$	65%	67%
connected in both steps	$\mathcal{E} \wedge \mathcal{A}$	20%	36%
for which no relations were discovered	$\neg(\mathcal{E} \vee \mathcal{A})$	11%	8%

Table 8.10: Senses connected with existing, $\mathcal{E}$, and superordinate $\mathcal{A}$ relations

Below we explain the reasons for the lack of relations for the group $\neg(\mathcal{E} \vee \mathcal{A})$. All these senses for which no relations were discovered neither from Figure 7.1: step 9 nor from Figure 7.1: step 14 were assigned to tags using statistical disambiguation. This means that their lexical neighbourhood has a high statistical relatedness to the tagspaces of the tags with which they were connected. However, there are no senses in the tagspace to which they can connect or their supersenses were filtered out during Figure 7.1: step 11. *Wood* is one such sense which was correctly assigned to the tag `wood` in the context of tagspace T={`sunset, dock, water, clouds, wood, upnorth, buelah, crystallake, michigan, canon, sigma1020mm, wide, wow`}.

Class: ***wood*** $\xrightarrow{subSenseOf}$ ***solid substance***
$\xrightarrow{superSenseOf}$ ***Birch***
$\xrightarrow{superSenseOf}$ ***Pine***
$\xrightarrow{superSenseOf}$ ***Beech***

In the tagspace of Dataset A (which is significantly smaller than the one of dataset B) there are no tags such as `birch`, `pine` and `beech`, which would have triggered the creation of the subordinate senses of *Wood*. Therefore, it does not relate to any existing senses. In addition, its supersense *solid substance* is filtered out because the terms `solid`, `substance` did not tag any resources is folksonomies. The same phenomenon is observed for the sense of tag `dress` in the contexts of T_1 = {`ragazza`, `abito`, `selfportrait`, `girl`, `dress`, `yellow`, `giallo`, `elisa`, `nothingdelicious`, `argh`, `partenzapercannesdel0207`, `consuddettoabito`, `maancheno`, `questoterribile`, `loscopriremosolovivendo`, `alloralocancello`, `ormailhaiscrittoelolasci`, `comandi`} and T_2 = {`red`, `fashion`, `dress`, `hat`, `reddress`, `redfashion`, `style`, `womensstyle`, `womensfashion`, `teenagefashion`, `teenagestyle`, `redandwhite`, `stylish`}.

Class: ***dress*** $\xrightarrow{subSenseOf}$ ***DurableGood***
$\xrightarrow{subSenseOf}$ ***EnvelopingCovering***
$\xrightarrow{superSenseOf}$ ***BallGown***
$\xrightarrow{superSenseOf}$ ***CoverUpDress***
$\xrightarrow{superSenseOf}$ ***BridalGown***
$\xrightarrow{superSenseOf}$ ***OffShoulderDress***
$\xrightarrow{superSenseOf}$ ***JumperDress***
"Dress is a specialization of Clothing [..] not drape down to the feet"

As a result, 89% of the correctly assigned senses were related to other senses in the structure while the 11% failure was caused by lack of tags defined with subordinate senses or due to filtering of supersenses. The repetition of the same experiment on the larger Dataset B is likely to return less disconnected senses and is a task for our future

work.

8.4 Evaluation of Tagspace Coverage

In this section we evaluate the tagspace coverage of FLOR-2 which is given by the percentage of tags that were assigned to correct senses. We measure two types of coverage, **total coverage** and **normalised coverage** M3.8. The total coverage is the ratio between the number of tags enriched correctly and the total number of tags. In Table 8.11 we see that the total number of tags enriched correctly with FLOR-1 was 281 and with the second is 994. We should point out that this value represents generic tags (because in the preliminary experiment we evaluated the enrichment in terms of generic tags). Therefore we calculated the approximate number of generic tags that were correctly enriched from FLOR-2 as 994 (the number of correctly enriched specific tags is 1421 and was measured using the sense assignment correctness process described in Section 8.2). As a result, the total coverage is calculated as $\frac{281}{2819} = 0.099$ for FLOR-1 and $\frac{994}{2819} = 0.33$ for FLOR-2, where 2819 is the number of generic tags in Dataset A.

	FLOR-1	**FLOR-2**
Correctly Enriched	281	994
Total Coverage	10%	33%
Normalised coverage	49%	81%

Table 8.11: Quantitative Improvement of the two versions of FLOR on Dataset A.

This is a significant improvement, yet the total coverage does not take into account the **vocabulary gap** between folksonomies and Knowledge Sources. This gap consists of the tags that were not enriched due to **folksonomic idiosyncrasies** or due to the lack of appropriate semantic entities that describe their meaning, namely the **sparseness of the Knowledge Sources**. As a result it does not indicate what percentage of the

non-coverage is caused due to knowledge sparseness and what percentage is caused by FLOR-2 failures (for example using a strict disambiguation in FLOR-1 restricted the number of covered tags).

Using the same rationale we presented in Section 4.6 we obtain the normalised coverage for FLOR-2. This is the ratio of tags that were enriched, $\mathcal{T}_{\text{A}}$ (see Section 3.6.1), compared to the tags that should be enriched but are not, $\mathcal{SE}$. The tags that should be enriched and are not can be described in terms of classic IR as "false negatives". We obtain an approximate number of false negatives as follows. We extract a random sample of the tags that were not enriched and the tags that were incorrectly enriched and try to enrich them manually. For each of these tags `t` we:

- locate the correct semantic entity in the available Knowledge Sources[4], and
- assess if the tagset of the non-enriched tag provides adequate information for the assignment of this entity (sense) to this tag.

If both conditions are met, `t` is added to the group of false negatives. The enriched tags $\mathcal{T}_{\text{A}}$ and the false negatives constitute the semantically covered tags M3.6, $\mathcal{T}_{\text{SS}} = \mathcal{T}_{\text{A}} + \mathcal{SE}$. We then calculate the normalised coverage, M3.8 as:

$$covn(\mathcal{T}, \mathcal{S}, FLOR-2) = \frac{\mathcal{T}_{\text{A}}}{\mathcal{T}_{\text{SS}}} = 81\%$$

In Section 4.6 we calculated the normalised coverage as 49% for FLOR-1 and using the method described above we calculate that the normalised coverage for FLOR-2 was 81%. This normalised coverage removed the bias given by the involvement of the Knowledge Sources and calculates the efficiency of the algorithm. Because the experiment with FLOR-1 was carried out in 2007 one may argue that since then there are more ontologies that provide adequate senses for the tags. Furthermore, the addition of WordNet as a Knowledge Source has indeed provided more appropriate senses for

[4]For the tags that were incorrectly enriched this is not necessary as they have been assigned already candidate senses

the tags. However, in the calculation of the normalised coverage the influence of these factors is eliminated because the false negatives are decided based on the existence of appropriate semantic entities in the Knowledge Sources.

Using the same process for Dataset B and FLOR-2 we obtained total coverage of 16% and a normalised coverage of 74%. We observe that the total coverage is quite low compared to the value we obtained for Dataset A. This is caused by the large number of lexical irregularities present in this dataset. We give more details on these in Section 8.5.1. Yet the large normalised coverage for Dataset B is close to the result we obtained for Dataset A. This justifies our decision to obtain a second measure that measures the performance of the algorithm and is independent of the vocabulary gap.

Including richer Knowledge Sources for the enrichment of folksonomies is part of our future work, yet we carried out a small experiment to understand the potential of such inclusion. We aimed to identify how additional resources can lexically cover the tags that were not covered by the current Knowledge Sources. We mapped the 84% of the tags from Dataset B that were not assigned a sense to DBpedia entities. For 87% of these unmapped tags we obtained at least one DBpedia entity. Yet this does not reflect the capability of DBpedia to semantically cover these tags in the context of the resource they appear. Therefore we repeated the evaluation process reported in Section 8.2 by selecting a random set of 100 assignments of tags to candidate entities and then assessing the appropriateness of one of these entities in terms of semantically describing the tags in their resource context. In 79 cases there was at least one DBpedia entity found to represent the meaning of the tag while for 21% of the tags none of the discovered DBpedia entities corresponded to their meaning in the tagset. Although this is a small experiment it already demonstrates that the inclusion of Linked Open Data resources can drastically improve the semantic coverage of tags.

8.5 Additional Analysis

FLOR-2 use a plethora of methods and heuristics during the enrichment process. The analysis presented in the previous sections focused on the final output of the algorithm. Yet, some of the results provided by the intermediate processes of FLOR-2 are also interesting and provide significant insights towards the improvement of the algorithm, and the nature of folksonomies and Knowledge Sources.

8.5.1 Lexical Isolation

In this section we provide a short analysis of the Lexical Isolation process, described in Section 7.3.1. From the 2819 tags of Dataset A, 1103 were removed, resulting in a tagspace of 2784 tag assignments and 1716 generic tags. This is a removal of 39% as opposed to the 59% isolated from FLOR-1 (see Section 4.6). In this instance, the 20% more tags kept was due to elimination of the WordNet filtering used to identify non-English tags in the preliminary experiment. In more detail, 3.5% of the isolated tags where shorter than 3, {`bw`, `jc`, `wc`}. 13.2% contained numbers, {`lovely1`, `save2`, `top10`} and 69.3% were isolated because they were infrequent (i.e., did not belong to clusters of frequently co- occurring tags). Finally 14% of the excluded tags were idiosyncratic.

Dataset B contains 69.099 generic and 223.537 specific tags. 153.394 (68%) of the tags passed the isolation phase, and for the 32% that were isolated, the distribution is as follows. 12.7% of the tags were ruled out by the idiosyncratic tag isolator, 5.2% contains numbers and 4.3% is shorter than 3. 8.9 % of the tags contain special characters and finally 68% of the tags were infrequent. It should be pointed out that the percentages of isolated tags may vary depending on which isolator was used first. For example the tag `mariasbirthday2009` is infrequent, idiodyncratic and contain numbers. However if the infrequent isolator is called first, this tag is classified under the infrequent tags.

In the next section we describe the overlap of vocabularies between the datasets and the Knowledge Sources.

8.5.2 Semantic Entity Discovery

In this section we discuss the process of matching a tag to a semantic entity either from online ontologies or WordNet and explore the vocabulary gap and the vocabulary overlap among tagspaces and Knowledge Sources.

The tags that passed through the lexical isolation, 61% of Dataset A and 68% of Dataset B, were used in the semantic entity discovery phase (Figure 7.1: step 3) to identify entities which may define their meaning. 59% of the generic tags from Dataset A and 47% of Dataset B were matched against at least one semantic entity. For the rest of the generic tags there was no semantic entity in any Knowledge Source that could match them. Table 8.12 depicts the percentages of non-covered tags for the two datasets. We obtained these numbers by evaluating a random sample of the non-covered tags.

Tag Type	Example	Dat. A	Dat. B
Non-English	`bleu`, `caer`, `abito`	31%	35%
Idiosyncratic	`allstars`, `catchycolors`, `cmwd`	17%	13%
Adjectives	`cool`, `lovely`, `alone`	15%	15%
Person Names	`julia`, `ahmed`, `deby`	3%	2%
Not Covered	`agip`, `bicyclette`, `chrysler`	15%	12%
Compound & Misspelled	`cityhall`, `bodylanguage`	19%	23%

Table 8.12: Tags for which FLOR-2 failed to identify candidate senses

The percentages shown in Table 8.12 are approximate values since they were acquired from a random sample of the uncovered tags. 31 - 35% of them are non-english, and as a result cannot be found either in online ontologies, whose majority is written in

English, or in WordNet. Idiosyncratic tags (17 - 13%), are also not likely to be found in online ontologies or WordNet. Adjectives (15%) and person names (3 - 2%) are out of the scope of FLOR-2 enrichment. This is because adjectives usually describe personal opinions [51] and are thus excluded from the entity search in WordNet. Personal names are underspecified in the tagspaces and not covered by online ontologies therefore are excluded from the enrichment process.

Also, 15 - 12% of tags could not be found in the online ontologies indexed by Watson, but are commonly used in folksonomies. For example, the tag **agip**, does not appear either in ontologies or in WordNet. Yet, it is frequent enough in the tagspace of Flickr, to have clusters of related tags:

C_1:	{ferrari, marlboro, fiat, mountain, formula1, nikon, motorsport}
C_2:	{sardinia, sardegna }
C_3:	{station, gas }
C_4:	{italy, distributore, italia, benzina, gasstation}

the majority of which indicate that "agip" is the brand name of a gas distributor. The case of brand names such as **agip** is a common case of non-coverage from FLOR-2 due to lack of semantic entities that define them in Knowledge Sources. Neither online ontologies nor WordNet cover these concepts, however, they are quite frequent in the tagspaces[5].

In addition to these, which are not covered lexically, there are other tags, which are covered lexically but they are not covered semantically, i.e., the correct sense is not available. This is the case of {converse, poi}, which exist in the Knowledge Sources but only with one of their meanings. Consider for example, the tag poi, which is encountered in the context of tagset T = {singapore, asia, night, movement, blur, fire, panorama, panoramic, handheld, poi} [6]. The only candidate sense discov-

[5]They are frequent enough to have clusters of related tags
[6]http://www.flickr.com/photos/arjunpurky/179834257/

ered for this tag is:

Sense: ***poi*** $\overset{subSenseOf}{\longrightarrow}$ ***dish*** $\overset{subSenseOf}{\longrightarrow}$ ***nutriment***
W(1) *"Hawaiian dish of taro root [...] often allowed to ferment"*

However, the correct sense for the tag in the context of T refers to the spinning fire game. Incidentally, one of the tag clusters extracted from Flickr for `poi` contains the tags {`fire, night, flame`} which overlap with T and could be used to disambiguate the sense of fire game. This is a frequent case where the information provided by folksonomies is adequate for the definition of a tag meaning but the respective knowledge is not available in any Knowledge Source. Yeung et. al [129] studied how different contexts of tags (identified statistically) overlap with WordNet. They also found that a large percentage of these contexts could not be mapped to WordNet. For the resolution of such issues further Knowledge Sources which contain more concepts should be investigated. This is part of our future work.

Finally the 19 - 23% of tags that was not mapped against a semantic entity are compound or misspelled tags that FLOR-2 failed to tokenise and correct in order to map to an entity. The larger value obtained for Dataset B, 23%, justifies its lower normalised coverage. In addition, the total number of non-covered tags in the entity discovery step is larger for B than A (53 versus 41%) which justifies the lower total coverage presented in Section 8.4.

In the following sections we analyse the execution and output from the next FLOR-2 processes using the results for Dataset A. Considering that sense assignment and tag coverage results for both Dataset A and B (which were obtained using different groups of evaluators) were similar we can assume that the outcomes of the following studies provided for Dataset A can approximately apply to Dataset B without loss of generality.

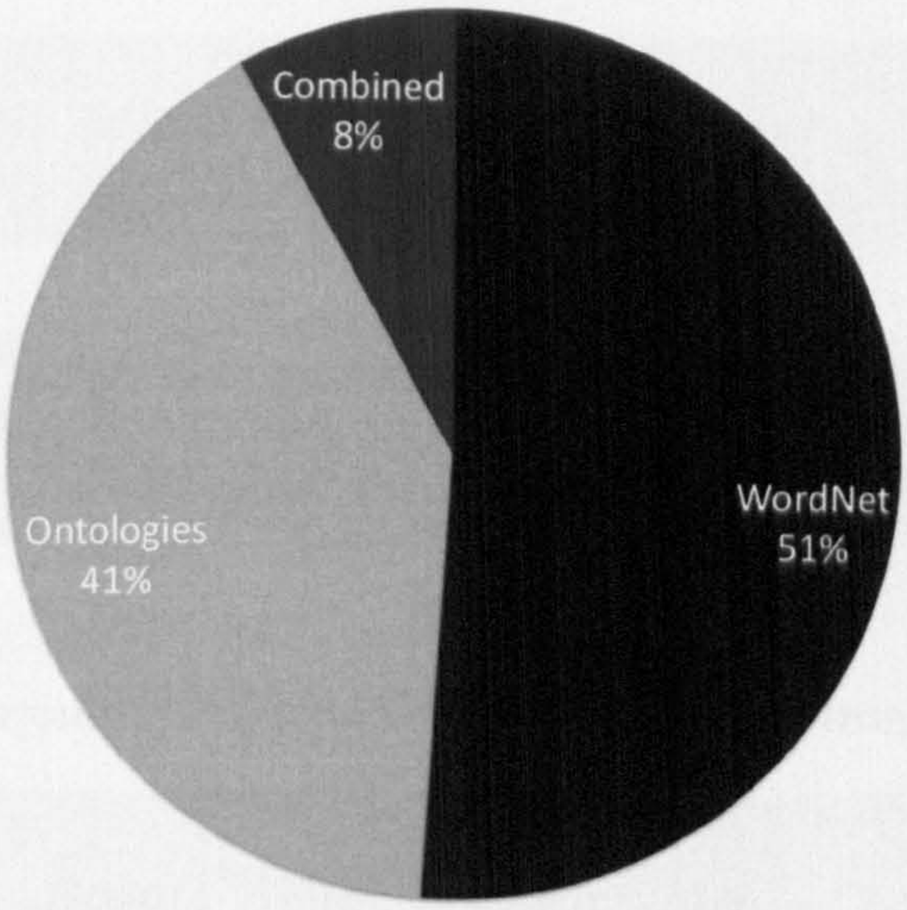


Figure 8.1: Provenance of all candidate senses assigned to the tags of DataSet A.

8.5.3 Sense Creation

The next step in the enrichment process, after the tags have been mapped to filtered semantic entities, is the exploitation of these entities for the creation of candidate senses. In this section we look into the statistics associated with sense provenance and present some indicative cases for sense creation.

For the tags of Dataset B the number of total candidate senses was 9672 and the approximate number of senses correctly assigned to the tags was 4105. For the tags of Dataset A, a total of 3760 candidate senses were created that potentially represent their meaning and 849 senses were correctly assigned to tags finally. This equals to a mean 3.7 candidate senses per tag. As shown in Figure 8.1, 51% of the candidate senses originated from WordNet synsets, 41% of them from ontological entities and only 8% were created using a combination of semantic entities from both Knowledge Sources.

Furthermore, 35% of the candidate senses were created from a single ontological entity and 48% from one WordNet synset. Examples of senses created from unique WordNet synsets, which do not share synonyms with other synsets are:

Synset: ***Maui, Maui Island*** $\rightarrow$ ***Island***
"the second largest of the Hawaiian Islands"

Synset: ***Parmesan*** $\rightarrow$ ***Cheese***
"hard dry sharp-flavored Italian Cheese"

Typical examples of senses that were created from a single ontological entity:

Class: ***Beauty*** $\overset{subClassOf}{\longrightarrow}$ ***Abstract Idea***
"the quality of being pleasing aesthetically, in an ideal sense"

The similarity of this entity with the WordNet synsets for beauty:

Synset: ***Beauty*** $\rightarrow$ ***Appearance***
"the qualities that give pleasure to the senses"

Synset: ***Beauty*** $\rightarrow$ ***Exemplar***
"an outstanding example of its kind"

is low due to the high heterogeneity in the definitions of the senses (parents) as well as their lexical information. Similarly, due to heterogeneity of definitions and low lexical relatedness the following senses failed to become one sense and remained separate senses, created from different semantic entities.

Class: ***Bridge*** $\overset{subClassOf}{\longrightarrow}$ ***Spot Features***
(Ontology 1)

Class: ***Bridge*** $\overset{subClassOf}{\longrightarrow}$ ***LandTransitway***
(Ontology 2) *"Bridge is the subclass of LandTransitways that are artifacts [...] over a natural surface"*

Class: ***Bridge*** $\overset{subClassOf}{\longrightarrow}$ ***SolidSurfacePathThroughAir***
(Ontology 3) $\overset{subClassOf}{\longrightarrow}$ ***Path-Simple***
"Bridges are elevated roadways, usually over water or some other pathway artifact"

Class: ***Bridge*** $\overset{subClassOf}{\longrightarrow}$ ***ManmadeOutdoorLocation***
(Ontology 4) *"Perhaps a bridge is not a location..."*[7]

This example demonstrates the phenomenon that many ontologies define entities tailored to certain tasks and valid in specific scopes, they are poor lexically or structurally (therefore do not provide enough evidence for sense integration) and their value on the enrichment is low. The entity filtering step (Figure 7.1: steps 4, 11) uses various heuristics to exclude such entities but further analysis is needed in order to improve the process of filtering out such noise. This is important for supporting the reuse of existing ontologies on the web.

In the following we discuss another phenomenon that affected sense integration, the inconsistently modelled knowledge in ontologies. Figures 8.2 and 8.3 display two semantic entities obtained when searching online ontologies for `party`.

These were transformed into the following senses:

Class: ***Party*** $\overset{subClassOf}{\longrightarrow}$ ***Actor***
(Ontology 1)

Class: ***Party*** $\overset{subClassOf}{\longrightarrow}$ ***Locatable***
$\overset{subClassOf}{\longleftarrow}$ ***Actor***
(Ontology 2)

We note that ontology 1[8] defines *Party* as a subclass of *Actor* and ontology 2[9] as a superclass. FLOR-2 does not deal with such types of contradictory knowledge and

[8] http://www.csl.sri.com/users/denker/owl-sec/ton/security_template.owl
[9] http://trajano.us.es/ isabel/EHR/Demographic_RM.owl

Details for http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#PARTY (view as graph)
Back

- In http://trajano.us.es/~isabel/EHR/Demographic_RM.owl
  - *Class*
  - **abstract**: true
  - **label**: ENTIDAD
  - **CEN**: healthcare agent
  - **HL7**: Entity
  - **label**: PARTY
  - **Purpose**: Ancestor of all party types, including real world entities and their roles. A party is any entity which can participate in an activity. The meaning attribute inherited from LOCATABLE is used to indicate the actual type of party
  - **subClassOf**: http://trajano.us.es/~isabel/EHR/Common_RM.owl#LOCATABLE
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#reverse_relationships: **domain**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#ROLE: **subClassOf**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#relationships: **domain**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#source: **range**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#identities: **domain**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#ACTOR: **subClassOf**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#target: **range**
  - http://trajano.us.es/~isabel/EHR/Demographic_RM.owl#contacts: **domain**

Figure 8.2: A semantic entity for *Party*

Details for http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#Party (view as graph)
Back

- In http://www.csl.sri.com/users/denker/owl-sec/ton/security_template.owl
  - *Class*
  - **subClassOf**: http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#Actor
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#isTrustedBy: **range**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#isTrustedBy: **domain**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#hasTrustedToken: **domain**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#employedBy: **range**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#trusts: **range**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#trusts: **domain**
  - http://kmi-web05.open.ac.uk:81/cache/1/8bb/5959/67443/f422524090/1a9258731283283ac#signedBy: **range**

Figure 8.3: Another semantic entity for *Party*

therefore these two senses are not merged despite the fact that they represent the same meaning.

Due to such failures of sense integration there are some cases where the tags are assigned a large number of candidate senses, which is disproportionate to their senses in reality (and leads to the large number of candidate senses 3.7). 30% of the tags are assigned only one sense, 18% two and 14% three candidate senses. The 38% is assigned more than three senses. This justifies our decision to perform sense ranking (Figure 7.1: step 7) prior to sense disambiguation.

8.5.4 Sense Disambiguation

In this section we present an overview of the correctly assigned senses in terms of the disambiguation method used (graph-based or statistical) and in terms of their provenance. In particular, we calculate the distribution of sense assignment correctness for graph-based disambiguated versus cluster disambiguated senses. In addition we estimate the ratio of correctness for senses originating from WordNet, ontologies and a combination of the two.

Figure 8.4 presents the distribution of correct, incorrect and undecided senses in three categories based on their provenance. As we can see in Figure 8.4, the majority of the correctly assigned senses (42%) has been created with a combination of semantic entities from WordNet and ontologies. This is an interesting outcome compared to the result of Figure 8.1, which shows that the minority of candidate senses is of mixed provenance. This also justifies our decisions on sense ranking based on their mixed provenance.

With regards to the performance of disambiguation methods, the graph-based disambiguation (relation) performed slightly better (51%) than statistical disambiguation (cluster) which accounts for the 49% of the correctly assigned senses. However, in

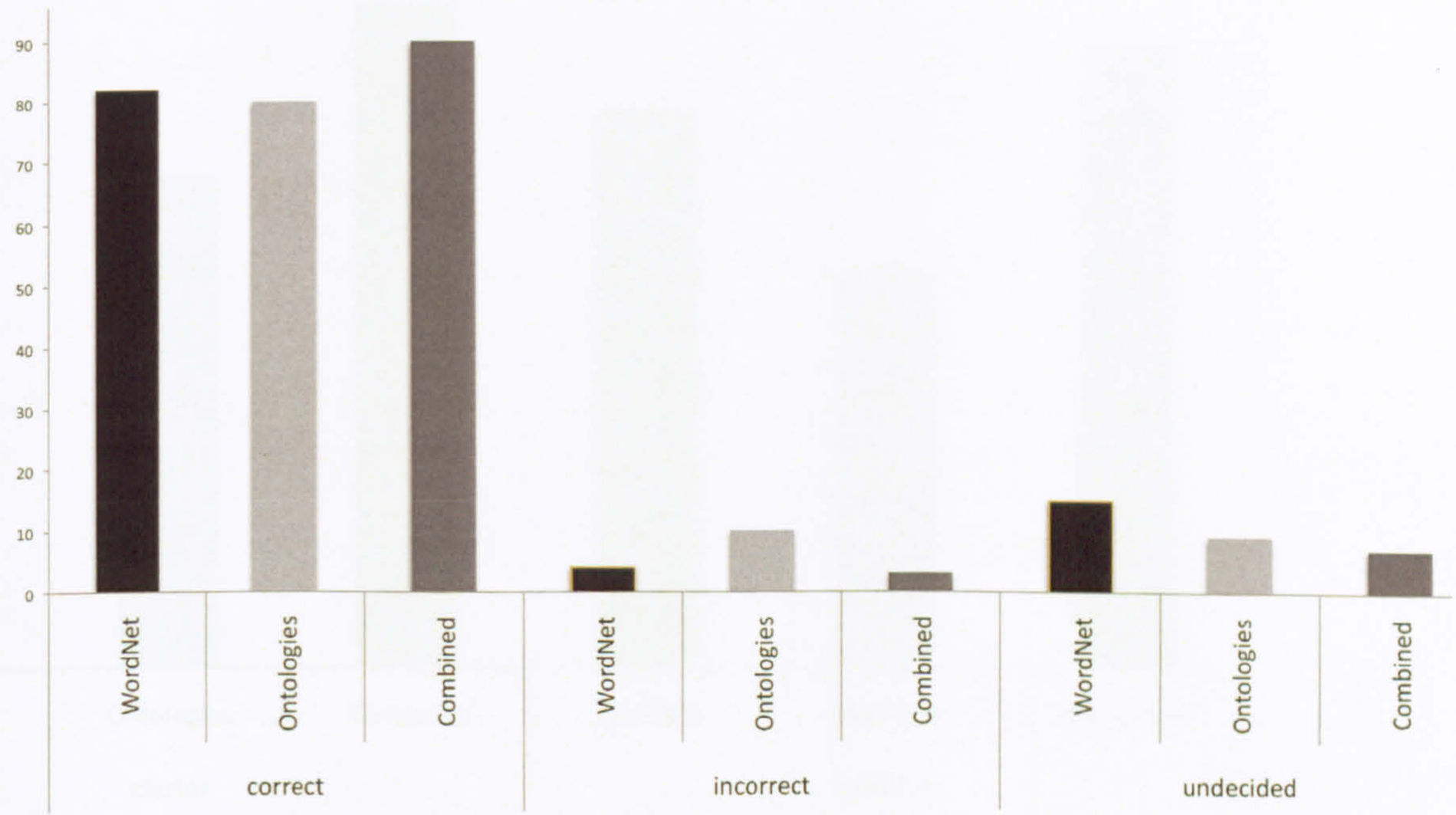


Figure 8.4: Sense Disambiguation Correctness by Sense Provenance

Figure 8.5: Sense Disambiguation Correctness by Disambiguation method

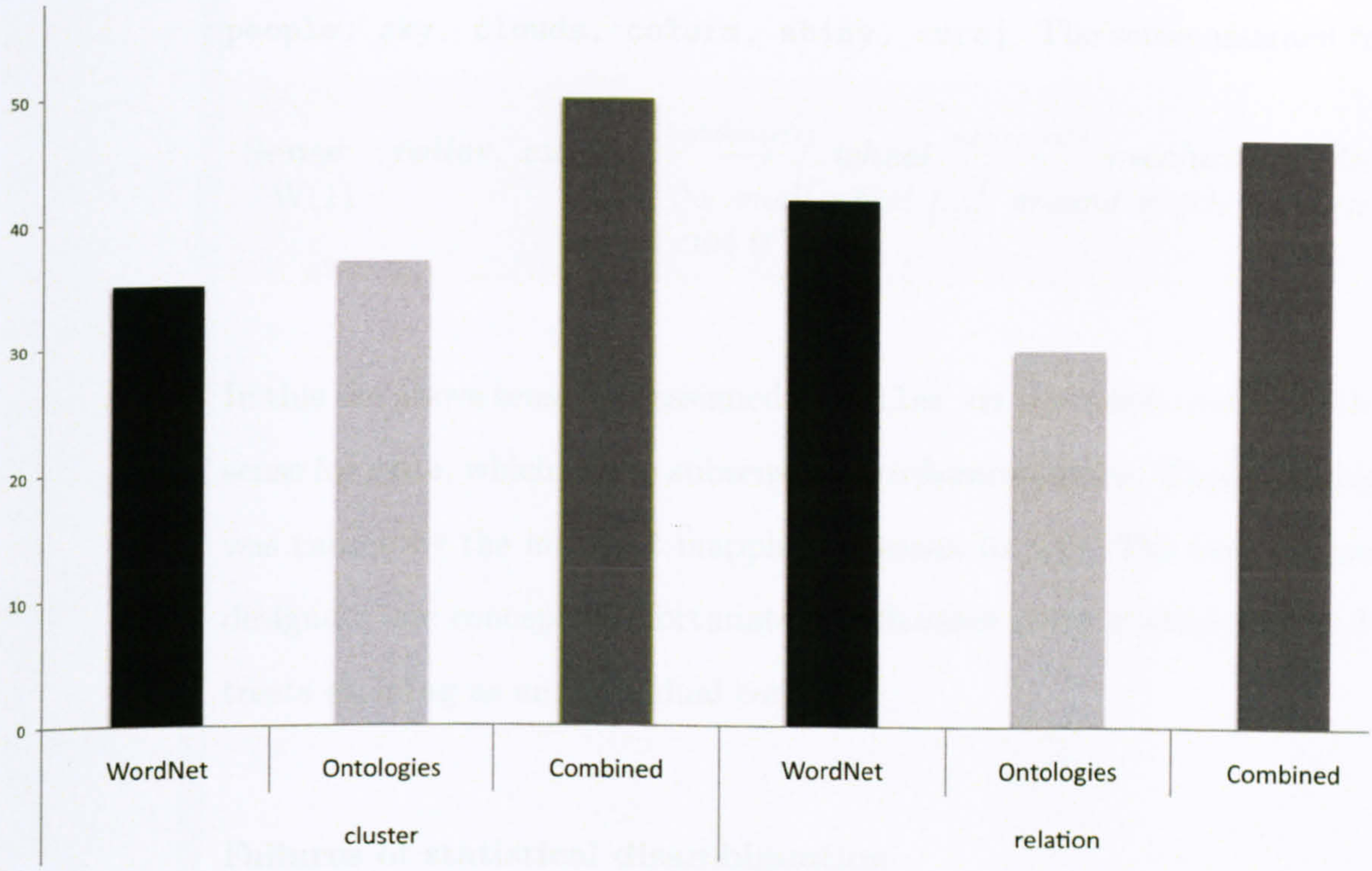


Figure 8.6: Provenance and disambiguation methods for the senses correctly assigned to the tags of DataSet A.

Figure 8.6 we observe that cluster base disambiguation performed slightly better with senses of combined provenance. This is explained as follows. The combined provenance senses have more diverse lexical neighbourhood, and the probability of higher relatedness with the tagset, even when formal relations are lacking, is higher. Overall this justifies our decision to incorporate a statistical disambiguation method in FLOR-2.

In the following we highlight some cases of disambiguation failures for graph-based and statistical disambiguation and highlight the case of inadequate tag context for disambiguation.

Failures of graph-based disambiguation

In this section we describe a case of graph-based disambiguation failure. Consider the tag `roller`, which belongs to the tagset T= {`ride, roller, coster, life, time, lifetime, scary, tall, white, red, brown, light, wow, yahoo, air, hands,`

`people, sky, clouds, colors, shiny, cure`}. The sense assigned to `roller` is:

Sense: ***roller, curler*** $\xrightarrow{subSenseOf}$ ***wheel*** $\xrightarrow{subSenseOf}$ ***mechanical device***
W(1) *"a small wheel [...] around which the hair is wound to curl it"*

In this the above sense was assigned to `roller` via the graph overlap with the candidate sense for `ride`, which is also subsense of *mechanical device*. This disambiguation failure was caused by the incorrect mapping of senses to tags. The tags `roller` and `coster`, designate one concept. Unfortunately such cases are not addressed by FLOR-2 which treats each tag as an individual concept.

Failures of statistical disambiguation

Statistical disambiguation exploits the collective relatedness of the tagset with a candidate sense's lexical neighbourhood using measure M7.7. However, measure M7.7 penalises the senses that have a larger lexical neighbourhood which does not overlap well with the tagset. For example consider the tag `artist` in the context of the tagset T= {`wizard, oz, art, studio, office, artist, collection, metal, desk`}. The sense:

Sense: ***artist*** $\xrightarrow{subSenseOf}$ ***agent***
O(1) *"A person or group or any agent who creates or performs art [...]"*

was assigned to the tag because of the higher overlap of its lexical neighbourhood compared to the neighbourhood of the sense:

Sense: ***artist*** $\xrightarrow{subSenseOf}$ ***creator*** $\xrightarrow{subSenseOf}$ ***participant***
O(1) $\xrightarrow{subSenseOf}$ ***creator*** $\xrightarrow{subSenseOf}$ ***person***
W(1) *"a person whose creative work shows sensitivity and imagination"*

Comparing the tags of T with the two lexical neighbourhoods {`agent`} and {`creator,`

participant, person} we observed that they overlap best with the terms **creator** and **agent**. However, the terms **participant** and **person** return a low relatedness with the tags of T and this reduces the overlap of the second sense with T. This disambiguation is not incorrect, because the first sense, which was assigned to **artist**, describes the meaning of the tag. However, this sense is not the optimal because it is not as descriptive as the second one.

Ambiguous language and underspecified tags

In some cases tags are not used with their literal meaning or the context is quite vague. An example of this was presented in Chapter 4, Figure 4.6 for the tag **volume**. A similar case to this was discovered for the image of the tag **beer** which was used to annotate the image of Figure 8.7[10]. The correct candidate sense was found for beer, but the tagset does not provide useful information either for graph-based disambiguation or for statistical disambiguation. The tagset is {**knit, knitting, beer, knit beer, project365, moments, 3082007**} and belongs to the groups "*Crazy for Knitting, The Knitting Club*", *CRAFT*", *tricot e crochet*", *Project 365!*", *365 Moments*", *Play Food!*". Such photos are taken in order to participate in a topic specific group of interest and therefore are cases for idiosyncratic tagging. In Section 7.3 we presented our heuristics for ruling out idiosyncratic tags they do not rule out idiosyncratic resources (this is the first instance we encountered such an example). The identification of idiosyncratic resources remains an issue of our future work.

Another example of poor context, consider the tag **alberta** which belongs to tagset T= {**trees, sunrise, sky, red, clouds, alberta**}. The only candidate sense for this tag refers to the canadian province. Although looking at the image and other textual information the user can infer that this is the correct sense for the tag, the contextual information given by its tagset is low and results to a failure of disambiguation with

[10] **http://www.flickr.com/photos/katknits/1000096206/**

Figure 8.7: An image of "knit beer"

either techniques. We should note that the evaluators of the sense assignment task in some cases requested for the actual resource tagged with the tagset. The tagset alone did not provide enough information for the judgements of correctness of a sense assignment.

In the following section we summarise the outcomes of this evaluation.

8.6 Summary

In this section we presented the evaluation of FLOR-2 with two datasets and obtained evidence on the performance improvement of the algorithm and the processes that need further improvement. In addition we acquired insights on folksonomies, the Knowledge Sources used for the enrichment, and our approach to automatically enriching folksonomies with existing semantics.

We evaluated FLOR-2 from three different perspectives, sense assignment (Section 8.2), semantic aggregation (Section 8.3) and tagspace coverage (Section 8.4) using two datasets of 250 and 25000 resources. For each of the two datasets FLOR-2 assigned correct senses to the 81% and 74% of the tags which were discovered in Knowledge Sources, with an approximate precision of 93%. The processes of entity discovery, sense integration and sense disambiguation account for the failures. Yet, the comparison of FLOR-2 to FLOR-1 showed a **significant improvement in the coverage of tags**, while the enrichment precision was maintained to the same levels. In terms of semantic aggregation FLOR-2 discovered relations among 89% of the senses that were correctly assigned to tags, while failure to identify relations was caused by the lack of overlap of the senses with the vocabulary of the tagspaces.

The low vocabulary overlap accounts also for the low total coverage of FLOR-2. Indeed only 33% of the tags from Dataset A and 16% from Dataset B were correctly enriched with senses. These are quite low percentages but are not caused by failures of the algorithm but by the following issues:

- There is a plethora of idiosyncratic and underspecified tags in folksonomies.
- These tag categories are difficult to distinguish and they are not covered by the Knowledge Sources.
- The Knowledge Sources are sparse even for tags that do not belong in the above categories.

In Sections 8.5.2, 8.5.3 and 8.5.4 we described how the combination of different Knowledge Sources and different disambiguation methods contributes to the correctness of sense assignment. We identified issues that need further analysis including:

- investigation of methods for the selection and evaluation of entities valuable to the enrichment process

- identification of additional context for underspecified tags
- resolution of conflicting knowledge

Still, despite these issues the performance of the algorithm was satisfactory. The major challenge we identified is the selection of additional Knowledge Sources which can complement the enrichment with entities that do not belong to ontologies and WordNet. These issues are part of our future work.

Chapter 9

exFLORe: Search on Enriched Tagspaces

In this chapter we describe exFLORe, an algorithm that exploits the semantic structures produced by FLOR-2 for the purposes of improving search in folksonomies. exFLORe translates the query keywords to senses and retrieves the resources associated with these senses.Finally, it presents the results in ranked groups.

9.1 Introduction

exFLORe is a query algorithm that makes use of the semantic structures created by FLOR-2 in order to improve search in folksonomies. Currently, folksonomy search is limited to matching search keywords against the tags (or other textual descriptions) of resources. For example, consider a query for resources related to European lakes phrased as {`europe lake`}. This can only return those resources that are explicitly tagged with both these keywords. However, other relevant resources might exist that are not tagged with exactly these keywords but rather with their semantic variations, for example {`italy, lake`} and {`balaton, hungary`}. This is an example of basic

level variation which, along with polysemy and synonymy, pose limitations on folksonomy search. Existing approaches have been proposed for the improvement of search in folksonomies (Section 2.3.2) either utilising semantics [72, 73, 95] or statistical approaches [17, 27, 85, 133] to address some of the phenomena of polysemy, synonymy and basic level variation and allow for additional functionalities such as result diversification [42]. Our approach addresses the underlying cause of these phenomena which is the lack of a semantic structure that can explicitly express the relations among {`italy`, `lake`, `europe`, `balaton`, `hungary`} and so on.

In the previous chapters we explained how FLOR-2 automatically structures the input tagspaces. In this chapter we introduce exFLORe, an algorithm that exploits this structure to improve folksonomy search. exFLORe supports traditional keyword-based querying, and does not introduce a new search paradigm. Its novelty lies in the translation of user keywords to senses, which are then used for the retrieval of relevant resources.

The approach used by exFLORe is different to other approaches that address the problem of Semantic Search on the web [46]. The latter make use of semantics as a means to expand the user queries, which are then compared against the textual annotations of the resources. Our approach exploits the fact that the resource space itself is annotated with a semantic structure. Therefore, rather that matching the user queries to the textual annotations of the resources (i.e., tags) they are matched against the resources' semantic descriptions (i.e., senses). In the following sections we present the details of exFLORe (Section 9.2) and exemplify its use on the scenario of folksonomy search (Section 9.3).

9.2 The exFLORe query algorithm

9.2.1 Approach

exFLORe exploits the semantic annotation of tagspaces created by FLOR. Consider the structure presented in Figure 9.1 which was created in accordance to the FLOR ontology (Section 3.5). At the bottom we observe the specific tags {`16668_Balaton`, `16668_Hungary`, `16668_Europe`} assigned to resource 16668. Each of them is linked to a sense via the relation *hasDefinition*. Although not explicitly, the resources are also linked to these senses in the following manner:

- a resource is related to a tag: 16668 $\overset{isTaggedWith}{\longrightarrow}$ `16668_Balaton`
- a tag is related to a sense: `16668_Balaton` $\overset{hasDefinition}{\longrightarrow}$ *Balaton*

Therefore, we can assume that the relation:

- 16668 $\overset{isTaggedWith}{\longrightarrow}$ `16668_Balaton` $\overset{hasDefinition}{\longrightarrow}$ *Balaton*

means that resource 16668 is connected to the sense *Balaton*, which is a richer annotation compared to the one provided by tag `16668_Balaton`.

Our approach is influenced by the work of Navigli et. al [89]. This approach exploits the semantic networks of senses as a means for query expansion. For each query keyword, they discover candidate senses (using WordNet and ontologies) and extracts their semantic networks, which are equivalent to our definition of semantic neighbourhoods. Then, the semantic networks of different senses are intersected in order to find connecting paths among candidate senses. A score is assigned to each connecting path. The senses which lead to the highest path score are selected and their lexical information (synonyms, hypernyms, related terms) is used to expand the query. The expanded query is then matched against the textual description of the resources.

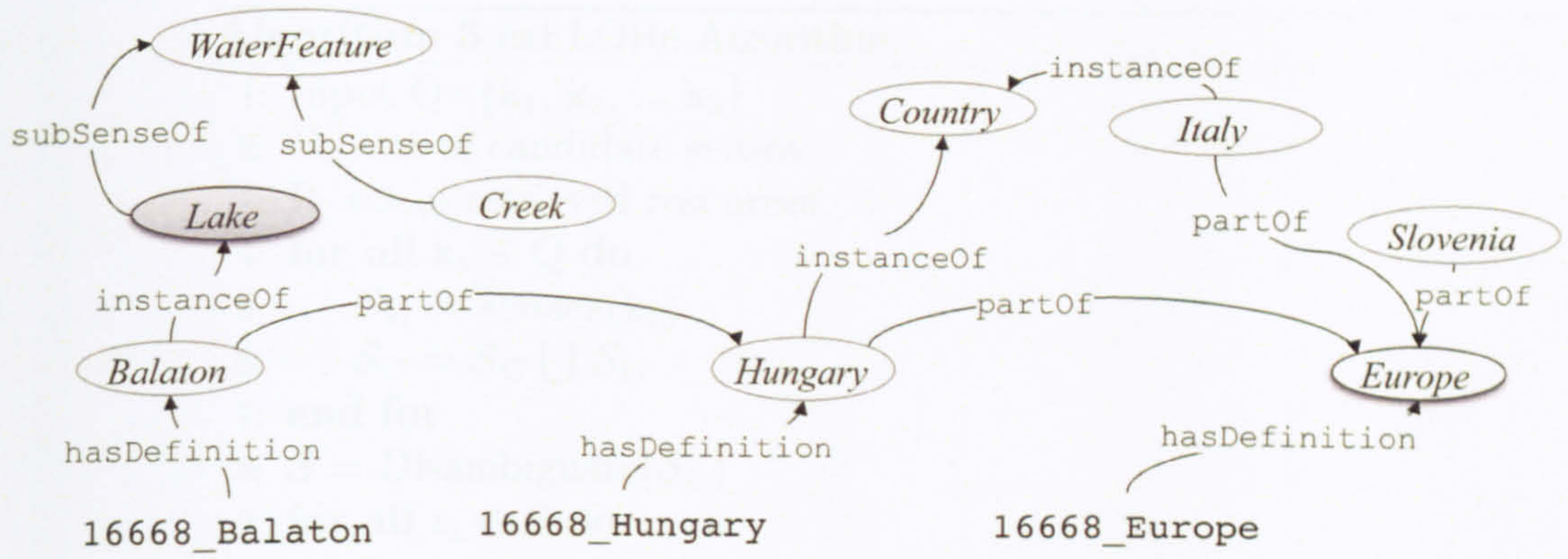


Figure 9.1: A structure of senses for the resource 16668

Our approach is similar because we also select the most appropriate senses using disambiguation that exploits their connecting path. However, we also use statistical disambiguation to cater for the lack of such paths. Once the appropriate senses are identified, instead of comparing their lexical information to the textual descriptions of the resources, we look for resources whose tags are defined by these senses.

9.2.2 Algorithm

The process of exFLORe is described in Algorithm 3. exFLORe takes as input a keyword query Q=$\{\mathtt{k}_1, \mathtt{k}_2, .., \mathtt{k}_N\}$ and for each keyword, $\mathtt{k}_i$, it locates its candidate senses, $\mathcal{S}_{\mathtt{k}_i}$, from the semantic structure of the tagspace (Algorithm 3: 4-7). The superset of candidate senses for all query keywords $\mathcal{S}_C$ is used in the same disambiguation process described in Section 7.5.2. Each keyword, $\mathtt{k}_i$ is considered as the tag to be disambiguated, t, and the set of keywords Q, contextualises $\mathtt{k}_i$ in the same manner that a tagset contextualises t. The disambiguation of $\mathcal{S}_C$ leads to a set of senses, $\mathcal{S}$, one for each query keyword (Algorithm 3: 8). For example, for Q=$\{$`lake`, `europe`$\}$ we obtain two senses defining their meaning, i.e., $\mathcal{S}$ = $\{$*Lake*, *Europe*$\}$.

Algorithm 3 exFLORe Algorithm

1: Input Q=$\{\mathtt{k}_1, \mathtt{k}_2, .., \mathtt{k}_N\}$
2: $\mathcal{S}_C$, set of candidate senses.
3: R, set of retrieved resources.
4: **for all** $\mathtt{k}_i \in$ Q **do**
5: $\quad \mathcal{S}_{\mathtt{k}_i} = senses(\mathtt{k}_i)$
6: $\quad \mathcal{S}_C = \mathcal{S}_C \bigcup \mathcal{S}_{\mathtt{k}_i}$
7: **end for**
8: $\mathcal{S}$ = Disambiguate($\mathcal{S}_C$)
9: **for all** $s_i \in \mathcal{S}$ **do**
10: $\quad \mathcal{S}_i = s_i \bigcup sub(s_i) \bigcup part(s_i) \bigcup ins(s_i)$
11: $\quad$ **for all** s $\in \mathcal{S}_i$ **do**
12: $\quad\quad$ R = R $\bigcup res(s)$
13: $\quad$ **end for**
14: **end for**
15: **for all** r $\in$ R **do**
16: $\quad$ **for all** $s_1 \in \mathcal{S}_1$, $s_2 \in \mathcal{S}_2$,.. $s_N \in \mathcal{S}_N$ **do**
17: $\quad\quad$ Create $S_{1,2,..,N}$
18: $\quad\quad$ **if** Overlap (r, $S_{1,2,..,N}$)=TRUE **then**
19: $\quad\quad\quad$ $S_{1,2,..,N}$: $R_A = \bigcup$ r
20: $\quad\quad$ **end if**
21: $\quad$ **end for**
22: **end for**
23: Organise($\{R_A, R_B,.., R_M\}$)
24: **if** $|\{R_A, R_B,.., R_M\}| < 4$ **then**
25: $\quad \mathcal{S}_i = s_i \bigcup sup(s_i) \bigcup type(s_i)$
26: $\quad$ Repeat Steps 11-23
27: **end if**
28: Rank($\{R_A, R_B,.., R_M\}$)
29: Output = ($\{R_A, R_B,.., R_M\}$)

For each of the senses in $\mathcal{S}$, s_i, a sense space $\mathcal{S}_i$ is created using its subordinate senses (subsenses, instances and parts of s_i) from the semantic structure (Algorithm 3: 9-10). For example:

$$\mathcal{S}_{\text{lake}} = \{lake,\ balaton,\ como,\ ..\} \text{ and}$$
$$\mathcal{S}_{\text{europe}} = \{europe,\ italy,\ hungary,\ ..\}$$

For each sense in the sense space of s_i, the resources that are connected with it are added in the result set R (Algorithm 3: 11-14). This step concludes the resource retrieval

process. All the resources that are connected with the senses of the query keywords, or their subordinate senses are included in R. The following steps of Algorithm 3 concern the categorisation and ranking of results.

The algorithm creates different combinations of senses (Algorithm 3: 16-17), $S_{1,2,..,N}$ using s_i from different $\mathcal{S}_i$'s as follows:

- {*europe*, *lake*}, {*europe*, *balaton*}, ...
- {*italy*, *lake*}, ...
- {*hungary*, *como*}, ..

Then, the overlap of each resource r $\in$ R with these combinations of senses is calculated (Algorithm 3: 18) by verifying that r is connected to the senses of the group. If the overlap is positive, then r is placed in R_A which contains the resources that overlap with $S_{1,2,..,N}$ (Algorithm 3: 19). This categorisation process, leads to multiple appearances of r in different groups, since it is quite likely for r to overlap with many groups of senses (for example, if resource r is connected with the senses of {*europe*, *lake*, *balaton*} it will appear in both resource groups {*europe*, *lake*} and {*europe*, *balaton*}).

To overcome this problem, the groups of resources are organised (Algorithm 3: 23), based on the following rule. Each resource that belongs to more than one group is maintained only in the set with the most specific senses. For example if r appears both in the groups representing {*europe*, *lake*} and {*europe*, *balaton*}, it will be removed from the group that represents {*europe*, *lake*}, because its concepts are more generic.

In case the groups of resources based on subordinate senses are less than 4 the superordinate senses, supersenses (if the sense originates from classes) and types (if the sense originates from instances) are extracted the process is repeated with these (Algorithm 3: 24-27). The reason we expand with superordinate if the subordinate groups

are less than 4 is because 4 is the average number of clusters returned by folksonomy search.

Finally, the resources and the groups are ranked based on the following.

- For two resource groups R_A and R_B, $rank(R_A) > rank(R_B)$ if:
  - $|R_A| > |R_B|$
  - $\sum_{r_a \in R_A} rank(r_a) > \sum_{r_b \in R_B} rank(r_b)$
- For two resources r_1 and r_2, $rank(r_1) > rank(r_2)$ if:
  - r_1 is connected to more senses than r_2
  - r_1 is connected to more specific senses r_2
  - r_1 contains more un-mapped query keywords than r_2. The un-mapped query keywords, are those that were not matched against a sense from the structure[1].

This algorithm addresses the phenomena of polysemy, synonymy and basic level variation in the following manner. First, it aligns the query keywords to senses. The senses are characterised by a set of synonyms and are used to define a set of tags, which lexically match some of these synonyms. Therefore, even if the query keyword does not explicitly match synonym tags, these tags (and the resources they tag) are considered because they are defined by the same sense as the query keyword. With this process exFLORe addresses the issue of synonymy. Second, the query keywords are disambiguated using the disambiguation algorithms of FLOR-2 aiming to identify the appropriate sense for each keyword. In this way exFLORe address issues caused by the polysemy of keywords. Finally, by expanding the keywords's senses with subordinate senses, it addresses the issue of basic level variation by returning resources which refer to subordinate concepts without being explicitly tagged with them.

[1]These were used to help disambiguate the other keywords in (Algorithm 3: 8), but they were not used for retrieving resources.

9.3 Using exFLORe to improve search

In this section we present a web search application powered by FLOR and exFLORe. We use Dataset B, used for the evaluation of the FLOR algorithm (Section 8.2.2). As mentioned before, exFLORe supports traditional keyword-based search, hence the front page of the application consists of a search box. The results are presented in ranked groups based on their combination of senses and keywords and for each group, a short explanation for its results is provided (addressing the outcomes of our search experiments L5.4, L6.3 and L6.4).

Figure 9.3 depicts the results of the system for the query "*europe lake*". While the exact matching of keywords to tags used by the conventional folksonomy search would not have retrieved any results on this dataset, our system did return relevant images. Screen A shows the entire result set for the query. The results are grouped according to the various ways in which they match the query. For example, the first group of images contains four resources that are tagged with `italy` and `lake` and are connected to the respective senses. The sense of *Lake* exists in the query and is related to the resources, however the sense of *Italy* is included in the expanded query since according to the semantic structure of Figure 9.1 *Italy* $\overset{partOf}{\longrightarrow}$ *Europe*. Similarly the second group of images, contains *lake* and *Austria* and *Austria* $\overset{partOf}{\longrightarrow}$ *Europe*. The third group contains images tagged with *Balaton*($\overset{instanceOf}{\longrightarrow}$ *Lake*), *Hungary* ($\overset{partOf}{\longrightarrow}$ *Europe* and *Europe*. The first group of results is ranked higher because it contains a larger number of resources. The second group was ranked higher than the third, first because it contains more senses related to the query (*Balaton*, *Hungary* and *Europe* while the other group represents only *Austria* and *Lake*) and second because it contains more specific senses (*Balaton* $\sqsubseteq$ *Lake* and *Hungary* $\sqsubseteq$ *Europe*). Finally the user can view a maximised version of a result and its tags (Screen B) by clicking on it.

For single keyword queries, for example, {`animal`}, where no disambiguation can take place, the system returns different groups of results for each different narrower sense

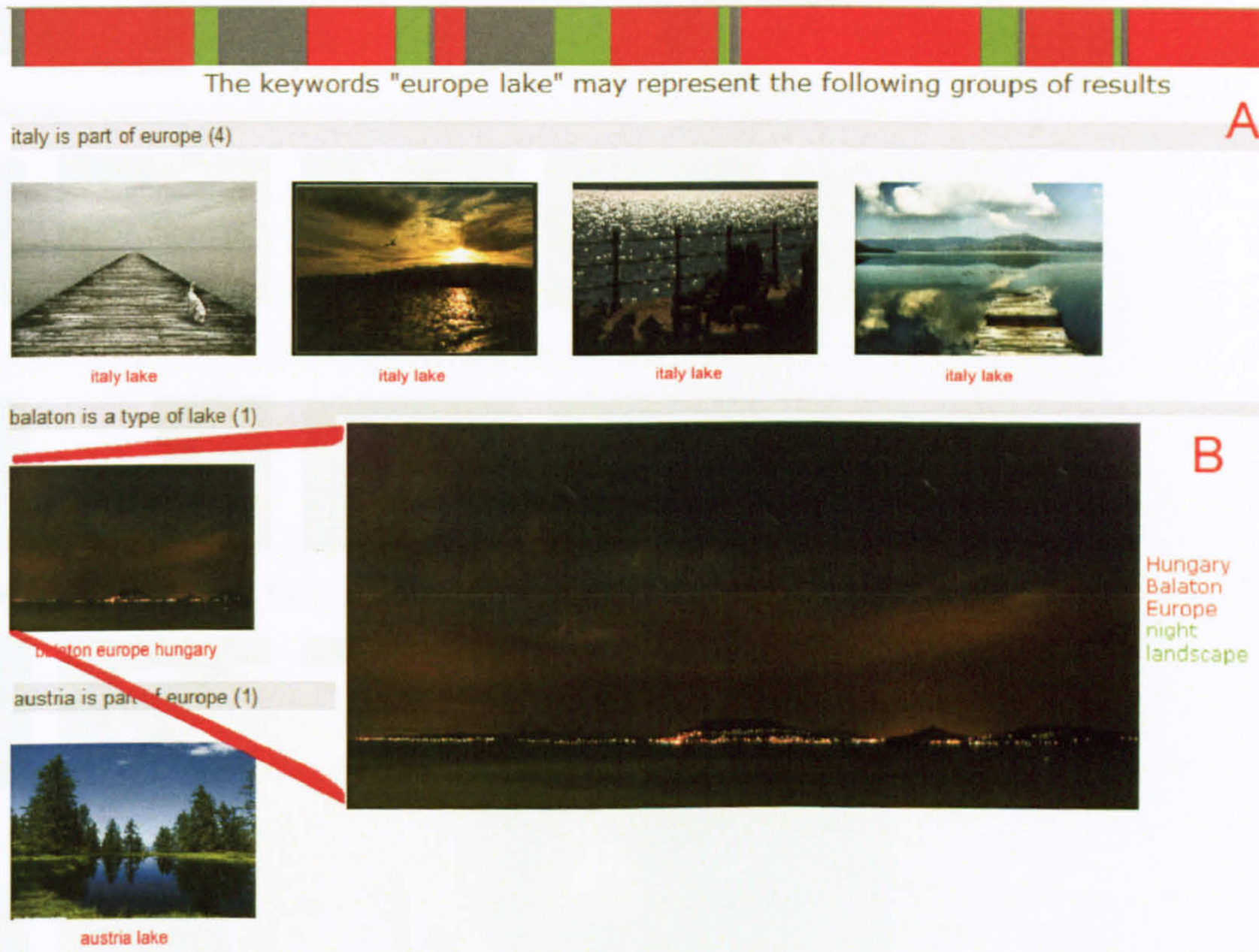


Figure 9.2: User Interface: Example results for the query Q={`lake`, `europe`}

of animal. Conventional folksonomy search cannot provide such result diversification.

9.4 Summary

The exFLORe algorithm exploits the structure of tagspaces created by FLOR to improve folksonomy search. It uses a controlled semantic space, which has been obtained using heterogeneous knowledge, i.e. online ontologies and WordNet. exFLORe translates the user keywords to senses and does not introduce a novel search paradigm.

Although we have not evaluated this approach in the scope of user experiment, we presented two example queries that indicate its improved effect on search compared to traditional keyword matching. The extensive user based evaluation of this approach including evaluation with the measures described in Section 3.6.2 is part of our future

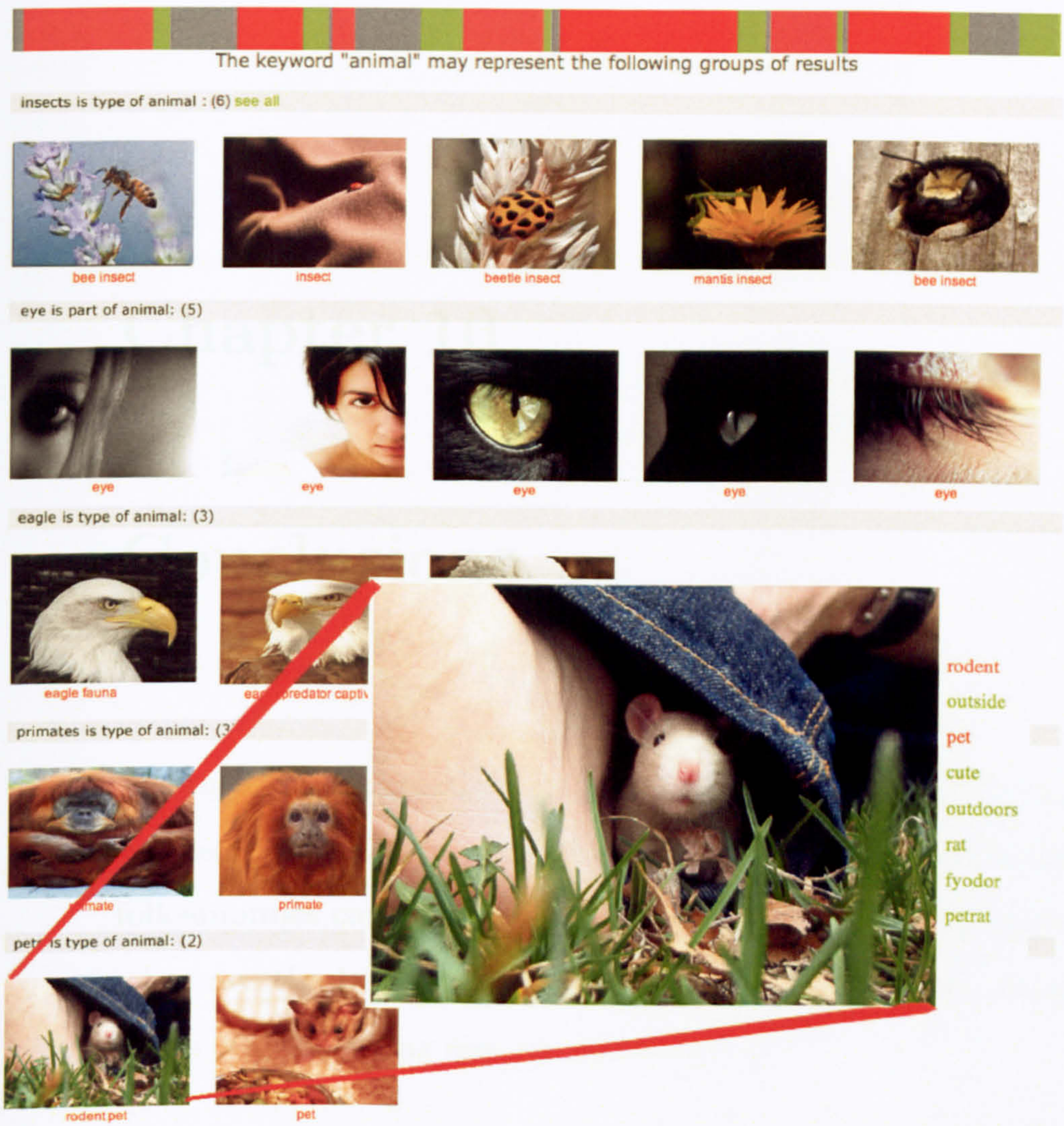


Figure 9.3: User Interface: Example results for the query Q={`animal`}

work.

Finally, the search offered by exFLORe is only one of the ways to exploit the structure created by FLOR-2 for content retrieval. In particular, the browsing capabilities of folksonomies are limited to the tag cloud navigation, which is a statistical summarisation of the tagspace. On the contrary, the structure provided by FLOR allows for a conceptual summarisation of the tagspace by exploiting the sense structure that annotates it.

ated? What measures and evaluation strategies should be established to quantify the performance of the enrichment methods?

RQ4: How can the enriched tagspaces be exploited and evaluated in the context of content retrieval? What methods should be created for improving folksonomy search utilising the enriched tagspaces? What measures should be established to asses the value of enriched tagspaces in search?

We have used an iterative approach and experimented with different Knowledge Sources, techniques and datasets in order to identify the most appropriate methods for automatic folksonomy enrichment. In particular, in chapters 4-6 we presented our initial approaches to enrichment and search and then built a folksonomy enrichment algorithm (Chapter 7) and a search mechanism (Chapter 9) that takes advantage of the semantic layer associated with folksonomies. Below we summarise the work and the outcomes presented in each chapter.

10.1 Summary of work

In **Chapter 2** we performed an analysis of the existing work on tagging systems from the perspectives of semantically structuring tagspaces and improving search. Although a plethora of approaches that improve search exist, the majority of them use statistical methods and address only some of the limitations posed by the phenomena of polysemy, synonymy and basic level variation. Approaches that structure folksonomies either require preselection of resources, or impose a different annotation process on the users, or they yield a generic hierarchical structure without explicitly specifying the relations among tags. Hence, we identified the need for an approach, which automatically applies structure on tagging systems by reusing existing semantics and without the need to preselect knowledge. Such an approach can explicitly specify the meaning of tags and their relations within a semantic structure.

In **Chapter 3** we described the problem at hand in more detail and defined the concepts of *resources*, *tagsets*, *clusters* and *tags* and their associations in folksonomies. We introduced the core objects of our approach, *Knowledge Sources*, *semantic entities* and *senses* and presented an ontology for the representation of the semantically enriched tagspaces. Finally, we introduced a set of measures for the evaluation of the semantically enriched tagspaces in terms of sense richness and tagspace coverage.

In **Chapter 4** we described the first version of the folksonomy enrichment algorithm FLOR-1 which disambiguates and semantically expands the tags using WordNet, and then enriches them with semantic entities from online ontologies. Evaluating FLOR-1 with a randomly selected dataset from Flickr we observed that using a hierarchical similarity measure on WordNet is too restrictive, given that tags in the tagsets tend not to be related with each other hierarchically. Therefore, the need to reconsider the usage ofWordNet as a disambiguation source emerged (L4.3). We also identified the need for alternative relatedness measures between tags (L4.1) and for the exploitation of statistical tag co-occurrence when semantic measures fail (L4.2).

In **Chapter 5** we studied the value of enriched tagspaces to the user in a search scenario. We applied FLOR-1 on a larger dataset from Flickr and used the enriched tagspace to perform query expansion using lexical variations, synonyms and hyponyms of the query keywords. Because of the low percentage of tags that was linked to ontological entities we conducted the experiment using only WordNet as a source of data. WordNet enabled satisfactory query expansion therefore we calculated that it can be considered as a Knowledge Source for enrichment (L5.3). We obtained the same outcomes of Chapter 4 (L4.1, L4.2) on the need for statistical relatedness measures for disambiguation of senses (L5.2). Aligning semantics with tags in a search scenario we *observed that the enrichment value* of semantic entities does not only depend on the richness of their specification in the Knowledge Source of *provenance but also on how* well their semantic neighbourhoods match the tagspace (L5.1). In terms of search we obtained useful insights on the user experience which included the fact that presentation

of results in groups is useful to the users and helps with query reformulation (L5.4)

Motivated by outcome (L5.3), in **Chapter 6** we performed a comparative study of WordNet and ontologies in terms of search and structure richness using the measures defined in Chapter 3. Our results showed that WordNet provides more synonyms for a sense compared to ontologies but ontological neighbourhoods of senses map better to the tagspace. Therefore a combination of the two Knowledge Sources would be beneficial for the enrichment of tagspaces (L6.1). We also compared the knowledge-based search algorithm with the cluster enabled search in folksonomies. We observed that statistically clustering the results returns fewer groups, caters for idiosyncratic tags but does not explain why a result belongs to a group (L6.2). On the contrary semantically-enabled search returns more meaningfully organised results with explanations but the number of groups should be restricted (L6.3).

Using the outcomes of the previous studies we created a set of requirements based on which we built the second version of the Folksonomy Enrichment Algorithm, FLOR-2, presented in **Chapter 7**. The improvements in FLOR-2 compared to FLOR-1 involve the processes of sense disambiguation (FLOR-2 uses a hybrid method that handle cases where no formal semantic relations between tags can be established), sense integration (the similarity for two senses exploits statistical co-occurrence of their lexical neighbourhoods) and semantic aggregation (FLOR-2 aggregates senses to a semantic structure, in contrast to FLOR-1). In addition, FLOR-2 exploits WordNet as a Knowledge Source for semantic entity discovery.

In **Chapter 8** we evaluated FLOR-2 in terms of sense assignment, coverage and relation discovery using two datasets. We used the results on the first dataset to compare the enrichment precision and average tag coverage obtained by FLOR-1 and FLOR-2. While the enrichment precision was maintained at the same levels as in FLOR-1, the normalised coverage significantly improved from 49% to 81%. The same precision and similar coverage values were obtained for the second dataset. A further analysis of the

senses that were correctly associated with tags, showed that the majority of correct assignments was obtained for senses of variable provenance (WordNet and ontologies) and for the combination of graph and statistical disambiguation. In sum, FLOR-2 achieved the same precision in sense assignment and a significantly larger normalised coverage than FLOR-1 (Chapter 4), where one resource (WordNet) and one method (hierarchical similarity) had been used to disambiguate the senses of tags. Furthermore, the FLOR-2 algorithm discovered relations for 89% of the senses in the tagspace, while the 11% that were not connected to any relations were senses whose neighbourhoods did not match the tagspace. This is an indication that the senses assigned to tags are valuable to the output semantic structure when their neighbour senses also cover the tagspace.

In **Chapter 9** we presented an algorithm that exploits the structure created by FLOR-2 and takes into account the outcomes of the search experiments (L5.4) (L6.3) and (L6.4). This algorithm aligns the query keywords to senses (addressing synonymy) and then applies the disambiguation methods used by FLOR-2 to decide the correct meaning of the query (addressing polysemy). Using subsumption and meronymy expansion (addressing basic level variation) the algorithm retrieves the relevant resources and presents them categorised in a meaningful way (result diversification).

10.2 Contributions

In this section we present the contributions of this thesis and discuss how they address the research questions.

Folksonomy Enrichment Algorithm. In Chapter 8 we presented FLOR-2, an algorithm that automatically applies semantics to tagspaces. FLOR-2 is domain independent, can be applied to any tagspace, does not require user feedback during the tagging activity (hence it does not interfere with the tagging activ-

ity), automatically selects the appropriate ontologies available in Watson, is able to handle heterogeneous knowledge, is unsupervised, and creates an explicit semantic structure supported by the FLOR ontology. The semantics of tags is discovered by using their resource context (tagset), semantic entities from the Knowledge Sources, and a combination of graph-based and statistical relatedness measures. FLOR-2 exploits the relations among semantic entities in Knowledge Sources to create relations between the senses that define the tags (RQ1: Part 1). The Knowledge Sources employed by the algorithm include all available ontologies in the Watson Semantic Web Repository. Appropriate ontological entities are selected, processed and reused for the creation of senses and relations for tags (RQ1: Part 2). The studies presented in chapters 4-6 highlighted the value of WordNet as a Knowledge Source for enrichment, therefore FLOR-2 includes WordNet in the employed Knowledge Sources. The evaluation performed in Chapter 9 confirmed the value of WordNet in the enrichment process (RQ2).

Evaluation measures for the semantic structures that represent the tagspaces. In Chapter 3 we introduced a set of measures for the evaluation of the semantic structures and used them in chapters 6 and 8 to asses the outputs of the enrichment algorithms. We described measures that evaluate the sense richness in terms of synonyms, subsenses and supersenses. In addition we provided a measure for the semantic coverage of tagspaces in semantic structures, and a measure for normalised coverage to asses the performance of FLOR-2. In Chapter 8 we presented an evaluation strategy for the correctness of sense assignments using random sampling and human evaluators and the measure of enrichment precision (RQ3). Finally we presented the measure of normalised increase which quantifies the percentage of results obtained using the enriched tagspaces compared to all the correct results (RQ4: Part 2).

Search Algorithm for enriched tagspaces. Finally, we presented a search algorithm that exploits the enriched tagspaces to improve search by addressing the

issues of polysemy, synonymy and basic level variation and at the same time allows for result diversification (RQ4: Part 1).

10.3 Outcomes and Future work

In this section we highlight the limitations of our approach, the characteristics of folksonomies and ontologies, as well as the lessons learnt from their combination.

10.3.1 Limitations of our approach and Extensions

The following are known limitations of our approach and are going to form part of our future work.

- In this thesis we focused on the ontologies available in the Watson Semantic Web Gateway and WordNet. This decision limited the lexical and semantic coverage of tags. As part of our future work, we plan to investigate additional ontology repositories and semantic resources, such as DBpedia, as well as exploit structured data sources, such as Freebase. In Section 8.4 we presented a small experiment that provided initial evidence on the significant improvement in semantic coverage of tags by such resources. Due to the modular architecture of FLOR-2 the integration of new Knowledge Sources with the existing ones depends on the process of sense creation (transformation of the semantic entities from the new Knowledge Sources to senses and transformation of the relations between entities to relations between senses). The rest of the processes operate on senses, therefore no significant alterations are needed in order to integrate additional Knowledge Sources. The integration of multilingual Knowledge Sources can improve the coverage of non English tags. Given that the only English source used by FLOR-2 is WordNet (during the process of sense integration in order to discover semantic

similarity among superordinate senses, see measure M7.3) and given that there are translated versions of WordNet in other languages[1], such an extension could be integrated in FLOR-2 with limited effort.

- The lexical isolation process of FLOR-2 rules out more than one third of the tags. We isolate different types of tags based on the assumption that they are less useful to the enrichment process because they are likely to represent other information than the content of the tag. For example, they may represent opinions of the users, membership in groups of interest, dates and symbolisms (tags with special characters). Although such tags are difficult to match against ontologies (there are no entities with the name of these tags e.g., *catchycolors*) they may be useful in the cases of underspecified tags or lack of adequate context for disambiguation. We plan to investigate the influence of idiosyncratic tags on the semantic enrichment further.

- Our approach is independent of the social interdependencies of tags and resources. Flickr is one folksonomy where in principle the tagsets of the resources are created by a single user[2]. We selected our evaluation datasets from Flickr and assume that the same performance of FLOR-2 can be achieved with additional folksonomies, such as Delicious or Last.fm, since the relations of tags and resources are the same across all tagging systems. Although this is a valid hypothesis we plan to evaluate FLOR-2 with data from other tagging systems.

- Another limitation of the work presented in this thesis is the lack of evaluation for exFLORe. Although empirical experimentations showed the value of FLOR-2 and exFLORe on search, we plan to perform an in depth evaluation with respect to the user experience in search, precision, normalised increase, and time performance.

- The lexical coverage of tags was also impeded by the failure of the lexical processing step to decompose compound tags (Table 8.12, compound tags). This is

[1] `http://www.illc.uva.nl/EuroWordNet/`

[2] Flickr allows for annotation of other users resources, however this functionality is not commonly used

also going to be part of our future work.

10.3.2 Characteristics of Knowledge Sources

One of the challenges addressed in this thesis was the reuse of ontologies in order to apply semantics to tagspaces. In the course of this study we identified a set of characteristics of ontologies that influence the performance of FLOR-2.

- Ontologies may include entities tailored to certain tasks and valid only in specific contexts, which from the point of view of folksonomy enrichment may provide low value. Such entities are poor either lexically or structurally (therefore do not provide enough evidence for sense integration) or their semantic neighbourhoods do not cover the tagspace (therefore they do not contribute relations). The entity filtering step of FLOR-2 (Section 7.4.2) uses a set of heuristics to exclude such entities but further analysis is needed in order to evaluate and select the most appropriate knowledge for reuse.

- Definitions of a concept across different ontologies can vary to a large degree. In Section 8.5.3 we presented an example of different definitions for *bridge* and we observed that the lexical and structural heterogeneity among the different senses of *bridge* caused a failure of merging and led to different senses for the same concept. FLOR-2 employs entity filtering and sense ranking to select the most appropriate senses for a given context. Yet, the existence of different senses referring to the same concept can lead to search problems (L6.4). Therefore, we plan to investigate the issue of sense integration further.

- Knowledge in ontologies can be defined in inconsistent ways and this was shown in Section 8.5.3 in the example of *party*, which different ontologies defined as a subclass or a superclass of *actor*. FLOR-2 does not attempt to resolve such conflicts and this issue will be part of our future work.

- Finally we observed some cases of non-coverage of tags by the Knowledge Sources. In particular, in some cases WordNet and ontologies did not cover tags either lexically (`agip`) or semantically (`poi`, `converse`). In the latter case the tags were mapped lexically but their intended senses in their resource contexts did not exist in the Knowledge Sources, leading to disambiguation failures. The introduction of richer Knowledge Sources is part of our future work as discussed in Section 10.3.1.

10.3.3 Characteristics of Folksonomies

- In this study we focused on the tags of resources rather than other lexical descriptions, such as titles and comments. We assumed that the resource context of a tag provides sufficient information for its disambiguation. This hypothesis applies to most of the cases, however, we encountered some examples where the resource context for a tag is vague or sparse. For example, during the evaluation of sense assignment correctness (Section 8.2) the judges commented that they were unable to make a decision due to the ambiguous tagset and requested to view the image that was tagged with it. In Section 8.5.4 we presented a case of disambiguation failure due to the sparse context of the tagspace for the tag `alberta`. Context expansion using statistical tag co-occurrence has been proposed in the literature [17] and we plan to investigate how it can improve the disambiguation process.

- An opposite issue to the one of sparse context was observed in the experiment presented in Chapter 5. The tagset of `beetle` contained three different and non related contexts (one related to cars, one related to plants and one related to Japan). FLOR-2 exploits the tagset context to perform tag disambiguation, therefore the existence of more than one contexts may cause disambiguation errors when disambiguating the tag with an incorrect context. As part of our future work, we plan to investigate further the influence of multi contextual tagsets on sense disambiguation.

- FLOR-2 assumes that each tag in the tagset represents one sense, yet a different case was encountered in Section 8.5.4 for the tags `roller` and `coaster`. These were used to describe the meaning of *roller coaster* but were split during the tagging process resulting to two separate tags/ concepts that do not reflect the intended meaning. The identification of senses from composite tags will also be an issue of our future work.

10.4 Outlook

In this thesis we presented our approach on automatically applying semantics to folksonomies and exploiting the enriched tagspaces in a search scenario. We combined the open ended and continuously evolving tagspaces of folksonomies with formal knowledge extracted from online ontologies and created semantic structures to represent the meaning and relations of tags. In this work we focused on semantically describing the meaning of tags that represent the content of the resources. Nevertheless, not all tags refer to resource topic. Specifically, tagspaces contain different types of descriptions such as dates, user information, places, user interests and more. The semantic description of such tags apart from content related tags can allow for a multifaceted organisation of the content and facilitate new intelligent retrieval applications. Although the strictly textual approaches (free tagging and keyword based search) are currently well established, initial efforts for the semantification of the content at annotation time have been adopted by the users in Web2.0. An early example of this is the usage of *hashtags* by Twitter[3] users. Hashtags are specially abbreviated names of the concepts and topics inherent in Twitter items and allow for the organisation and unambiguous retrieval of items referring to a particular concept or topic. Although this is not a fully fledged semantic approach, it shows that intelligent yet subtle diversions from the text-based paradigms are well received by the users. Additional semantically-enabled

[3] `http://twitter.com`

approaches, such as Semantic MediaWiki[4] and RDFa[5] already enable the generation of semantic content to a certain extent. Nevertheless, it is not realistic to believe that users will embrace paradigms that involve processes of laborious semantic annotation. Therefore, methods and algorithms such as the ones presented in this thesis that can exploit semantics, as well as unstructured content, are needed.

[4] http://semantic-mediawiki.org/
[5] http://www.w3.org/TR/xhtml-rdfa-primer/

Appendix A

Glossary

Tagspace	a set of resources $\mathcal{R}$=$\{R_1, .., R_{\|\mathcal{R}\|}\}$ and a set of tags $\mathcal{T}$ =$\{t_1, .., t_{\|\mathcal{T}\|}\}$
Specific Tag	a tag t that belongs to a resource R and is annotated with the resource name, i.e., R_t.
Generic Tag	if no reference is made to the instance of a tag t with respect to a specific resource then it is generic.
Tag Context (Resource)	the set of tags assigned to the same resource as the tag.
Tag Context (Cluster)	the set of tags globally associated with the tag in all entities of the tagspace (either resources or users).
Knowledge Source	a body of knowledge that contains semantic descriptions of concepts and explicitly defined semantic relations between them.
Semantic Entity	a Knowledge Source object that contains information that defines one concept.
Sense	an object that defines the meaning of a tag.
Semantic Neighbourhood	a set of the explicitly related concepts of a semantic entity or sense.

Lexical Neighbourhood	a set of lexical terms of the explicitly related concepts of a semantic entity or sense.
Lexical Coverage	the ratio of tags that are lexically covered by entities of a knowledge source.
Semantic Coverage	the ratio of tags that are semantically covered by entities of a knowledge source (when the meaning of tags is explicitly described by the entities of the knowledge source).
Normalised Coverage	(with respect to an enrichment algorithm) the ratio of tags correctly associated to entities of a knowledge source with respect to the tags that are semantically covered by this knowledge source.

Appendix B

FLOR Ontology

Class: TaggedResource

This class represents all folksonomy resources that are tagged with at least one tag.

$$\forall \, \mathrm{R} \in \mathcal{R} \; : \; |T_{\mathrm{R}}| > 0 \; \exists \; (TaggedRecource_R)^I \in (TaggedRecource)^C$$

This means that for each resource R in the set of resources $\mathcal{R}$ which is tagged with at least one tag (thus the cardinality of its tagset is not zero $|T_{\mathrm{R}}| > 0$) there exists one individual $(TaggedRecource_R)^I$ of the class $(TaggedRecource)^C$ that represents the tagged resource. In the example of Figure 3.8 $\mathrm{Resource_X}$ is an instantiation of the class *TaggedResource*.

Class: SpecificTag

This class represents all tags `R_t` $\in$ tagspace $\mathcal{T}$ that belong to the tagset of the resource R.

$$\forall \, \mathtt{R\_t} \in \mathcal{T} \; \exists \; (SpecificTag_{R\_t})^I \in (SpecificTag)^C$$

In other words, every tag R_t that annotates a resource R is represented by an instance of the class $(SpecificTag)^C$. This class represents the specific occurrence of the tag in the tagspace of a single resource. For example, in Figure 3.8 there is a tag **X_Europe** that belongs to the resource X and is denoted with the resource identifier. One instance of $(SpecificTag)^C$ is created for each occurrence of the tag **europe** in each resource.

Property:isTaggedWith

This is an object property that formalises the relation "*A tagged resource R is tagged with tag R_t*". Therefore, the domain of this property is $(TaggedRecource)^C$ and the range is $(SpecificTag)^C$. The cardinality of this relation is One-to-Many in order to represent the fact that one resource can be tagged with many tags, but each of these specific tags is only assigned to this specific resource. In the example of Figure 3.8 the following hold:

$$(TaggedRecource_X)^I \overset{isTaggedWith}{\longrightarrow} (SpecificTag_{X\_Hungary})^I$$
$$(TaggedRecource_X)^I \overset{isTaggedWith}{\longrightarrow} (SpecificTag_{X\_Balaton})^I$$

and $(SpecificTag_{X\_Europe})^I$, $(SpecificTag_{X\_Hungary})^I$, $(SpecificTag_{X\_Balaton})^I$ can only tag the resource X. To this end, $(SpecificTag_{Y\_Balaton})^I$ is assigned to resource Y and so on.

Class:Sense

This class represents the sense (concept or meaning) that is assigned to a specific tag in accordance to Definition 6 from Section 3.4. Each individual of this class has the properties described in the following.

Synonyms This is a set of words that denote the meaning of the specific individual.

For example, the synonyms of $(Sense_{Europe})^I$ could be {**europe**, **europa**, **evropi**}.

The Synonyms are associated with the sense via a Datatype property, defined in the ontology, the property **hasSynonym**.

$$(Sense_{Europe})^I \xrightarrow{hasSynonym} \texttt{europe}$$
$$(Sense_{Europe})^I \xrightarrow{hasSynonym} \texttt{europa}$$

Glosses Glosses is a set of natural language descriptions of the meaning of $(Sense_{Europe})^I$.

The glosses are also associated with a sense via a Datatype property **hasGloss**.

$$(Sense_{Europe})^I \xrightarrow{hasGloss} \textit{"the 2nd smallest continent"}$$

Semantic Entities These are the semantic entities discovered in the Knowledge Sources where this sense was initially defined and from where it was extracted. Such semantic entities can be ontological entities or WordNet synsets (Definition 5, Section 3.4) where a concept or entity is defined. Each semantic entity is associated with the sense via the property **isFoundIn**.

$$(Sense_{Europe})^I \xrightarrow{isFoundIn} http://ontology1.europa$$
$$(Sense_{Europe})^I \xrightarrow{isFoundIn} WordNet.synset.europe$$

Property: hasDefinition

This object property represents the relation between a specific tag and a sense that describes its meaning.

$$\forall\, \texttt{R\_t} \in \mathcal{T}\, \exists\, (Sense_A)^I \in (Sense)^C \;:\; Dfn(\texttt{R\_t}, Sense_A) = 1.$$

Its domain is $(SpecificTag)^C$ and its range $(Sense)^C$. The cardinality of this relation is Many-to-One, i.e. many specific tags may have the same sense, but the meaning of a specific tag in the context of the particular tagged resource is uniquely defined by one sense.

$$(SpecificTag_{X\_Hungary})^I \xrightarrow{hasDefinition} (Sense_{Hungary})^I$$
$$(SpecificTag_{Y\_Hungary})^I \xrightarrow{hasDefinition} (Sense_{Hungary})^I$$

This property, the $(Sense)^C$ and the properties described in Section B models the output of FLOR which is the explicit representation of tag meaning and the relations of the tags (via the relations of their senses). This property and the property *isFoundIn* allow for the implicit link of a tag to a semantic entity, in line with the existing ontologies [63, 98] as follows:

$$(SpecificTag_{X\_Europe})^I \xrightarrow{hasDefinition} (Sense_{Europe})^I \xrightarrow{isFoundIn} http://ontology1.europa$$

Sense Relations Properties

The final set of properties specified in the ontology are abstractly depicted in Figure 3.7 with a property named **relation**. This set of properties have both domain and range the $(Sense)^C$. In the following sections we briefly describe the most popular relations that connect the various senses and were discovered in the Knowledge Sources.

Property:subSenseOf This property is used to represent more specific (subordinate) senses of a sense. This is a broad relation used to represent *rdfs:subClassOf* [34] and WordNet *hyponym*.

$$(Sense_{Country})^I \xrightarrow{subSenseOf} (Sense_{Region})^I$$

Property:superSenseOf This property, is the inverse of *subSenseOf*, and represents the inverse of *rdfs:subClassOf* and WordNet *hypernym*.

$$(Sense_{Region})^I \xrightarrow{superSenseOf} (Sense_{Country})^I$$

Property:instanceOf This property is used to represent *rdf:type* and WordNet *instance.*

$$(Sense_{Hungary})^I \xrightarrow{instanceOf} (Sense_{Country})^I$$

Property:hasInstance This property, inverse of *instanceOf*, represents the inverse of *rdf:type* and WordNet *has instance.*

$$(Sense_{Country})^I \xrightarrow{hasInstance} (Sense_{Hungary})^I$$

Property:hasPart This property is a super-property for all the WordNet relations for meronymy such as substance meronyms, part meronyms and member meronyms.

$$(Sense_{Europe})^I \xrightarrow{hasPart} (Sense_{Italy})^I$$

Property:isPartOf This property is the inverse relation of *hasPart* represents the holonym relations form WordNet.

$$(Sense_{Hungary})^I \xrightarrow{isPartOf} (Sense_{Europe})^I$$

Appendix C

Sample of Sense Assignments from Dataset B

sa	tag	tagset	sense vector	JB1	JB2	JB3	JB4	JB5
1	Bridge	[Sunrise, Middlesbrough, E_500, Urbex, Transporter, Bridge, Long_exposure, ISO_200, 14_42mm, HDR]	[a stable, supported artifact, Structure, bridge, Building, construction, bailey_bridge, baileybridge, cantilever_bridge, cantileverbridge, cattle_guard, cattleguard, cattle_grid, ... truss_bridge, trussbridge, viaduct, span, support]	1	1	1	1	1
2	leaves	[torleyvideo, seeds, forest, melon, cinematic, hand, grendel_s, held, leaves, solid, sepia, dark, flea, shaky, shoots, camera, metal, children, vine, film, tone, amateur, hideo, kojima, look, watermelon, bussy, gear, meandering]	[AboveGroundPlantParts, leaves]	1	1	?	1	1
3	seeds	[torleyvideo, seeds, forest, melon, cinematic, hand, grendel_s, held, leaves, solid, sepia, dark, flea, shaky, shoots, camera, metal, children, vine, film, tone, amateur, hideo, kojima, look, watermelon, bussy, gear, meandering]	[AboveGroundPlantParts, seeds, plant parts, PlantPart, seed, EdibleLegumeSeed, Coconut-TheSeed, CocoaBean-TheSeed, SesameSeed, MustardSeed, Cashew-TheSeed, PlantDerivative, PlantGrain]	1	1	1	1	1
4	hawk	[vancouverisland, britishcolumbia, ferruginoushawk, raptor, hawk]	[AccipitrineFamily, BirdOfPrey, PuertoRicanBroadWingedHawk, HawaiianHawk, PuertoRicanSharpShinnedHawk, Osprey, Falcon, hawk, bird_of_prey, raptor, raptorial_bird, raptorialbird, eyas, tiercel, tercel, tercelet, goshawk, accipiter_gentilis,.....sea_eagle, seaeagle, pandion_haliaetus, pandionhaliaetus]	1	1	1	1	1
5	infrared	[trees, toboggan, b_w, woods, sled, forest, sledding, criticism_welcome, infrared, preserve, IR, muddyboots, outdoor, colorized, false_color, run, IL, hinsdale, bemis, illinoise, illinois]	[actinic_radiation, actinicradiation, actinic_ray, actinicray, infrared, infrared_light, infraredlight, infrared_radiation, infraredradiation, infrared_emission, infraredemission]	1	1	0	1	1
6	work	[work, granny, fun, afghan, colorful, bliss, progress, crochet, hexagon, wip]	[activity, wash, washing, lavation, action, job, operation, procedure, service, shining, polishing, heavy_lifting, heavylifting, housewifery, housework, housekeeping, ironing, busywork, make-work, logging, loose_end, looseend, unfinished_business, unfinishedbusiness,.. spadework, timework, undertaking, project, task, coursework, work]	1	1	?	1	?
7	old	[sit, strobe, dusty, high_contrast, wheelchair, abandoned, fire_training_college_urbex, old]	[age, old]	1	1	1	?	?
8	Old	[T, Blow, Old, T_shirt]	[AgeGroup, old]	1	?	1	?	?
9	grupo	[town, noite, centro, grupo, clicksp, sky, urbanistassp, augusta, city, sao_paulo, sunrise, sp, flickr, sampa, sampaist, night, ceu, urbanistas]	[Agent, a_class_of_agents_, aclassofagents, group, Group of People, GroupOfPeople, ACTOR, grupo, Village, People, picavanovanowan, Each_village, Eachvillage]	1	1	?	?	1
10	Oregon	[Water_Fountain, OSCON_2007, Portland_OR, OSCON, Oregon, Pidgeon, Bird, Portland, Fountain]	[american_state, americanstate, oregon, beaver_state, beaverstate]	1	1	1	1	1
11	california	[usa, berkeley, ucberkeley, california]	[AmericanState, california, american_state, colorado_desert, coloradodesert, golden_state, goldenstate, calif_, calif]	1	1	1	1	1
12	Florida	[umbrella, St_Petersburg, USA, beach, sky, Florida, Nikon, d80, Webel, orange, blue]	[AmericanState, florida, american_state, sunshine_state, sunshinestate, everglade_state, evergladestate]	1	1	1	1	0

13	illinois	[trees, toboggan, b_w, woods, sled, forest, sledding, criticism_welcome, infrared, preserve, IR, muddyboots, outdoor, colorized, false_color, run, IL, hinsdale, bemis, illinoise, illinois]	[AmericanState, illinois, american_state, prairie_state, prairiestate, land_of_lincoln, landoflincoln]	1	1	1	1	1
14	texas	[tower, texas, reunion, dallas]	[AmericanState, texas, american_state, lone_star_state, lonestarstate]	1	1	1	1	1
15	Flower	[VR, 55_200mm, Flower, Flora, Yellow, D40x, Nikon]	[angiosperm, flowering_plant, floweringplant, bloomer, peony, paeony, lesser_celandine, lessercelandine, pilewort, ranunculus_ficaria, ranunculusficaria, pheasant's-eye, adonis_annua, adonisannua, anemone, windflower, rue_anemone, ... french_honeysuckle, frenchhoneysuckle, centranthus_ruber, centranthusruber, flower]	1	1	1	1	1
16	flower	[tucheng, flower, bokeh, taiwan, sigma30mm1_4]		1	1	1	1	1
17	flower	[trees, garden, flowers, Strauch, Master_Photos, Baum, Blumen, Garten, tree, flower, rosa, Blume, pink, B_ume]		1	1	1	1	1
18	lac	[water, lagon, verte, carri_re, gypsum, underground, r_flexion, gypse, quarry, eau, souterraine, mirror, green, miroir, lac, vert]	[animal_product, animalproduct, garnet_lac, garnetlac, gum-lac, shellac, stick_lac, sticklac, seed_lac, seedlac, lac]	1	?	1	?	0
19	quarry	[water, lagon, verte, carri_re, gypsum, underground, r_flexion, gypse, quarry, eau, souterraine, mirror, green, miroir, lac, vert]	[animal, animate_being, animatebeing, beast, brute, creature, fauna, prey, quarry]	1	0	1	?	1
20	bear	[urban, futuristic, bear, portrait, forrest, surreal]	[animal, bear, CarnivoreOrder, .. ice_bear, icebear, polar_bear, ursus_maritimus, ursusmaritimus, thalarctos_maritimus, thalarctosmaritimus, sloth_bear, slothbear, melursus_ursinus, melursusursinus, ursus_ursinus, ursusursinus]	1	1	1	1	1
21	bird	[waves, wellington, islandbay, night, newzealand, bird]	[Animal, Pet Bird, PetBird, bird, OviparousAnimal, Nightingale, BlackGuillemot, KauaiAkialoa, Raven, CanadaGoose, Hummingbird, Vulture, Crane-Bird, PigeonGuillemot, ... VultureBird, warm-blooded animals, vertebrates, oviparous animals, terrestrial organisms, oriole, magpie, quail, condor, duck-animal, eagle, "An antarctic-dwelling non-flying bird", "any bird of prey", raptor, grebe, loon, albatross, falcon, lark, turkey, goose]	1	1	1	1	1
22	Bird	[Water_Fountain, OSCON_2007, Portland_OR, OSCON, Oregon, Pidgeon, Bird, Portland, Fountain]		1	1	1	1	1
23	Birds	[spring, finches, nature, babies, KansasCity, park, Missouri, LoosePark, Birds]		1	1	1	1	1
24	horse	[uniform, horse, pez, the_public_garden, ernie, police, boston, mounted]	[Animal, Pet Horse, PetHorse, horse, Livestock, HoofedMammal, EquineAnimal, HerdAnimal, Horse-WarmBlood, Horse-Stallion, Pony, Horse-Gelding, Colt, Horse-Domesticated, Horse-ColdBlood, DomesticAnimal, ...highstepper, chestnut, liver_chestnut, liverchestnut, bay, sorrel, palomino, pinto, equus_caballus, equuscaballus]	1	1	1	1	1
25	Architecture	[Silouette, Art, Peter_Zumthor, Swiss, Bregenz, Architecture, Gallery, Zumthor, Switzerland, Cube, Facade, Glass, Museum, Tree, Frosted]	[art, architecture, culture_and_art, cultureandart]	1	1	0	1	1
26	chair	[square_format, speaker, 120, SpittinShells, Japan, Tokyo, ptsix_TL, film, expired, 6_6, chair, medium_format]	[Artifact, Metal-Chair, Wooden-Chair, chair, Seat, Furniture, seating-furniture, stool, high-chair, armchair, wheelchair, chaise-longue]	1	?	1	1	1
27	video	[sundown, TC_80N3, residential, Canon_EOS_30D, time_lapse, neighborhood, street, Tijuana, sunset, lights, night, Tamron_17_50mm_f_2_8, video]	[Artifact, painting, decorative-artifact, picture, a_painted_picture, apaintedpicture, graphic_art, graphicart, abstraction, cityscape, daub, distemper, finger-painting, icon, ikon, landscape, miniature, illumination, monochrome, ... watercolour, visual_communication, visualcommunication, video]	1	0	0	1	0
28	minimalist	[van, Nikon, minimalism, England, London, minimal, minimalist, green, fairly_meaningless]	[artist, creative_person, creativeperson, minimalist]	1	1	1	1	0
29	van	[van, Nikon, minimalism, England, London, minimal, minimalist, green, fairly_meaningless]	[artistic_movement, artisticmovement, art_movement, artmovement, avant_garde, avantgarde, vanguard, van, new_wave, newwave]	0	0	0	?	0

Bibliography

[1] Amazon: Online shopping. http://www.amazon.com.

[2] Citeulike.org: Everyone's library. http://www.citeulike.org.

[3] Commontag. http://www.commontag.org/.

[4] Connotea: Organise share discover. http://www.connotea.org.

[5] Cross language evaluation forum. http://www.clef-campaign.org/.

[6] Dbpedia. http://dbpedia.org.

[7] Delicious: Social bookmarking. http://delicious.com.

[8] Flickr: Photo sharing. http://flickr.com.

[9] The friend-of-a-friend project. http://www.foaf-project.org/.

[10] Image retrieval in clef. http://imageclef.org/2010.

[11] Last.fm: The social music revolution. http://www.last.fm/.

[12] Linking open data. http://www.w3.org/DesignIssues/LinkedData.html.

[13] Ontology alignment evaluation initiative. http://oaei.ontologymatching.org/.

[14] Wikipedia: The free encyclopedia. http://www.wikipedia.org/.

[15] Youtube: Video broadcasting. http://youtube.com.

[16] Rabeeh Abbasi, Stefan Staab, and Philipp Cimiano. Organizing resources on tagging systems using T-ORG. In *Proceedings of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 97–110, Innsbruck, Austria, 2007.

[17] Rabeeh Abbasi and Steffen Staab. RichVSM: enRiched vector space models for folksonomies. In *Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 219–228, Torino, Italy, 2009. ACM.

[18] Hend S. Al-Khalifa and Hugh C. Davis. Measuring the semantic value of folksonomies. In *Innovations in Information Technology, 2006*, pages 1–5, 2006.

[19] Hend S. Al-Khalifa and Hugh C. Davis. Towards better understanding of folksonomic patterns. In *HT '07: Proceedings of the 18th conference on Hypertext and hypermedia*, pages 163–166, New York, NY, USA, 2007. ACM.

[20] Harith Alani, Wendy Hall, Kieron O'Hara, Nigel Shadbolt, Martin Szomszor, and Peter Chandler. Building a pragmatic semantic web. *IEEE Intelligent Systems*, 23:61–68, 2008.

[21] Sofia Angeletou, Marta Sabou, Lucia Specia, and Enrico Motta. Bridging the gap between folksonomies and the semantic web: An experience report. In *Proceedings of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 30–43, Innsbruck, Austria, 2007.

[22] Ruba Awawdeh and Terry Anderson. Improved search in Tag-Based systems. In *Proceedings of the 2009 Ninth International Conference on Intelligent Systems Design and Applications*, pages 288–293. IEEE Computer Society, 2009.

[23] Ricardo Baeza-Yates. From capturing semantics to semantic search: a virtuous cycle. In *ESWC'08: Proceedings of the 5th European semantic web conference on The semantic web*, pages 1–2, Berlin, Heidelberg, 2008. Springer-Verlag.

[24] Ricardo Baeza-Yates, Massimiliano Ciaramita, Peter Mika, and Hugo Zaragoza. Towards semantic search. In *NLDB '08: Proceedings of the 13th international conference on Natural Language and Information Systems*, pages 4–11, Berlin, Heidelberg, 2008. Springer-Verlag.

[25] Shenghua Bao, Guirong Xue, Xiaoyuan Wu, Yong Yu, Ben Fei, and Zhong Su. Optimizing web search using social annotations. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 501–510, New York, NY, USA, 2007. ACM Press.

[26] Grigory Begelman, Philipp Keller, and Frank Smadja. Automated tag clustering: Improving search and exploration in the tag space. In *Proceedings of the 15th International Conference on World Wide Web (WWW'06)*, 2006.

[27] Matthias Bender, Tom Crecelius, Mouna Kacimi, Sebastian Michel, Thomas Neumann, Josiane Xavier Parreira, Ralf Schenkel, and Gerhard Weikum. Exploiting social relations for query expansion and result ranking. In *ICDE Workshop (IEEE 24th International Conference on Data Engineering)*, pages 501–506. IEEE Computer Society, 2008.

[28] Richard Benjamins, John Davies, Ricardo Baeza-Yates, Peter Mika, Hugo Zaragoza, Mark Greaves, Jose Manuel Gomez-Perez, Jesus Contreras, John Domingue, and Dieter Fensel. Near-term prospects for semantic technologies. *IEEE Intelligent Systems*, 23:76–88, 2008.

[29] Tim Berners-Lee, James Hendler, and Ora Lassila. The semantic web, May 2001.

[30] Kerstin Bischoff, Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. Can all tags be used for search? In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 193–202, New York, NY, USA, 2008. ACM.

[31] Christian Bizer, Jens Lehmann, Georgi Kobilarov, Sören Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. Dbpedia - a crystallization point

for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3):154–165, September 2009.

[32] Uldis Bojars, John. Breslin, Vasilieos Peristeras, Giovanni Tummarello, and Stefan Decker. Interlinking the social web with semantics. *IEEE Intelligent Systems*, 23:29–40, 2008.

[33] Ron Brachman. Emerging sciences of the internet: Some new opportunities. In *Proceedings of the 4th European Semantic Web Conference, pages 1–3, Innsbruck,* Austria, June 2007.

[34] Dan Brickley and R. V Guha. Rdf vocabulary description language 1.0: Rdf schema, February 2004.

[35] Francesca Carmagnola, Federica Cena, Omar Cortassa, Cristina Gena, and Ilaria Torre. Towards a tag-based user model: How can user model benefit from tags? In *UM2007, User Modeling: Proceedings of the Eleventh International Conference*, volume 4511 of *Lecture Notes in Computer Science*, pages 445–449, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[36] Ciro Cattuto, Dominic Benz, Andreas Hotho, and Gerd Stumme. Semantic grounding of tag relatedness in social bookmarking systems. In *Proceedings of the 7th International Semantic Web Conference*, pages 615–631, Berlin, Heidelberg, 2008. Springer-Verlag.

[37] Rudi Cilibrasi and Paul Vitanyi. The google similarity distance. *Transactions on Knowledge and Data Engineering, IEEE*, 19(3):370–383, 2007.

[38] Tanguy Coenen, Dirk Kenis, Céline Van Damme, and Eiblin Matthys. Knowledge sharing over social networking systems: Architecture, usage patterns and their application. In *On the Move to Meaningful Internet Systems 2006: OTM 2006 Workshops*, pages 189–198, 2006.

[39] Céline Van Damme, Martin Hepp, and Katharina Siorpaes. Folksontology: An integrated approach for turning folksonomies into ontologies. In *Proceedings of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, 2007.

[40] Mathieu d'Aquin, Enrico Motta, Marta Sabou, Sofia Angeletou, Laurian Gridinoc, Vanessa Lopez, and Davide Guidi. Toward a new generation of semantic web applications. *Intelligent Systems, IEEE*, 23(3):20–28, 2008.

[41] Mathieu d'Aquin, Marta Sabou, Martin Dzbor, Claudio Baldassarre, Laurian Gridinoc, Sofia Angeletou, and Enrico Motta. Watson: A gateway for the semantic web. In *Proceedings of the 4th European Semantic Web Conference*, Innsbruck, Austria, 2007.

[42] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[43] Javier Diaz, Keyun Hu, and Melanie Tory. An exploratory study of tag-based visual interfaces for searching folksonomies. In *Proceedings of the 2009 British Computer Society Conference on Human-Computer Interaction*, pages 410–417, Cambridge, United Kingdom, 2009. British Computer Society.

[44] Takeharu Eda, Masatoshi Yoshikawa, Toshio Uchiyama, and Tadasu Uchiyama. The effectiveness of latent semantic analysis for building up a bottom-up taxonomy from folksonomy tags. *World Wide Web*, 12(4):421–440, December 2009.

[45] Christiane Fellbaum. *WordNet - An Electronic Lexical Database.* MIT Press, 1998.

[46] Miriam Fernandez, Vanessa Lopez, Marta Sabou, Victoria Uren, David Vallet, Enrico Motta, and Pablo Castells. Semantic search meets the web. *IEEE Semantic Computing*, 0:253–260, 2008.

[47] Claudiu S. Firan, Wolfgang Nejdl, and Raluca Paiu. The benefit of using tag-based profiles. In *LA-WEB '07: Proceedings of the 2007 Latin American Web Conference*, pages 32–41, Washington, DC, USA, 2007. IEEE Computer Society.

[48] Wai-Tat Fu, Thomas G. Kannampallil, and Ruogu Kang. Facilitating exploratory search by model-based navigational cues. In *Proceeding of the 14th international conference on Intelligent user interfaces*, pages 199–208, Hong Kong, China, 2010. ACM.

[49] Andres Garcia, Martin Szomszor, Harith Alani, and Oscar Corcho. Preliminary results in tag disambiguation using dbpedia. In *Knowledge Capture (K-Cap'09) - First International Workshop on Collective Knowledge Capturing and Representation - CKCaR'09*, September 2009.

[50] Domenico Gendarmi and Filippo Lanubile. Community-driven ontology evolution based on folksonomies. In *Community Informatics 2006*, France, 2006.

[51] Scott Golder and Bernardo Huberman. Usage patterns of collaborative tagging systems. *Journal of Information Science*, 32:198–208, 2006.

[52] Mark Greaves. Semantic Web 2.0. *IEEE Intelligent Systems*, 22(2):94–96, 2007.

[53] Tom Gruber. A translation approach to portable ontology specifications. *Knowl. Acquis.*, 5(2):199–220, 1993.

[54] Tom Gruber. Ontology of folksonomy: A mash-up of apples and oranges, 2005.

[55] Marieke Guy and Emma Tonkin. Folksonomies - tidying up tags?, January 2006.

[56] Tony Hammond, Timo Hannay, Ben Lund, and Joanna Scott. Social bookmarking tools (i): A general review. D-Lib Magazine, 2005.

[57] James Hendler. The dark side of the semantic web. *IEEE Intelligent Systems*, 22(1):2–4, 2007.

[58] James Hendler and Jennifer Golbeck. Metcalfe's law, web 2.0, and the semantic web. *Web Semantics: Science, Services and Agents on the World Wide Web*, 6(1):14–20, February 2008.

[59] Paul Heymann and Hector Garcia-Molina. Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical report, Stanford University, 2006.

[60] Andreas Hotho, Robert Jäschke, Cristoph Schmitz, and Gerd Stumme. Information retrieval in folksonomies: Search and ranking. In *Proceedings of the 3rd ESWC*, pages 411–426, Budva, Montenegro, 2006.

[61] Mark Huiskes and Michael Lew. *The MIR Flickr Retrieval Evaluation.* ACM, New York, NY, USA, 2008.

[62] Hak Lae Kim, Suk Hyung Hwang, and Hong Gee Kim. Fca-based approach for mining contextualized folksonomy. In *Proceedings of the 2007 ACM symposium on Applied computing*, pages 1340–1345, Seoul, Korea, 2007. ACM.

[63] Hak Lae Kim, Alexandre Passant, John Breslin, Simon Scerri, and Stefan Decker. Review and alignment of tag ontologies for semantically-linked data in collaborative tagging spaces. In *Semantic Computing, 2008 IEEE International Conference on*, pages 315–322, 2008.

[64] Hak-Lae Kim, Sung-Kwon Yang, John G. Breslin, and Hong-Gee Kim. Simple algorithms for representing tag frequencies in the SCOT exporter. In *IAT '07: Proceedings of the 2007 IEEE/WIC/ACM International Conference on Intelligent Agent Technology*, pages 536–539, Washington, DC, USA, 2007. IEEE Computer Society.

[65] Hong-Gee Kim, Suk-Hyung Hwang, Yu-Kyung Kang, Hak-Lae Kim, and Hae-Sool Yang. An agent environment for contextualizing folksonomies in a triadic context. In *Proceedings of the 1st International KES Symposium on Agent and*

Multi-Agent Systems - Technologies and Applications, pages 728–737, Wroclaw, Poland, 2007.

[66] Margaret E. I. Kipp. Exploring the context of user, creator and intermediate tagging. In *ASIS&T 2006 Information Architecture Summit*, March 2006.

[67] Georgi Kobilarov, Tom Scott, Yves Raimond, Silver Oliver, Chris Sizemore, Michael Smethurst, Christian Bizer, and Robert Lee. Media meets semantic web – how the BBC uses DBpedia and Linked Data to make connections. In *The Semantic Web: Research and Applications*, pages 723–737. 2009. 10.1007/978-3-642-02121-3_53.

[68] Christian Körner, Dominik Benz, Andreas Hotho, Markus Strohmaier, and Gerd Stumme. Stop thinking, start tagging: tag semantics emerge from collaborative verbosity. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 521–530, New York, NY, USA, 2010. ACM.

[69] David Laniado, Davide Eynard, and Marco Colombetti. Using WordNet to turn a folksonomy into a hierarchy of concepts. In *Proceedings of Semantic Web Application and Perspectives - 4th Italian Semantic Web Workshop*, pages 192–201, Bari, Italy, 2007.

[70] Ora Lassila and James Hendler. Embracing "Web 3.0". *Internet Computing, IEEE*, 11(3):90–93, 2007.

[71] Faith Lawrence and Schraefel. Bringing communities to the semantic web and the semantic web to communities. In *www06*, pages 153–162. ACM, 2006.

[72] Sun-Sook Lee and Hwan-Seung Yong. Component based approach to handle synonym and polysemy in folksonomy. In *Computer and Information Technology, 2007. CIT 2007. 7th IEEE International Conference on*, pages 200–205, 2007.

[73] Sun-Sook Lee and Hwan-Seung Yong. Tagplus: A retrieval system using synonym tag in folksonomy. In *Int. Conf. on Multimedia and Ubiquitous Engineering*, pages 294–298, Seoul, Korea, 2007.

[74] Vladimir Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[75] Freddy Limpens, Fabien Gandon, and Michel Buffa. Collaborative semantic structuring of folksonomies. In *WI-IAT '09: Proceedings of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence and Intelligent Agent Technology*, pages 132–135, Washington, DC, USA, 2009. IEEE Computer Society.

[76] Huairen Lin, Joseph Davis, and Ying Zhou. An integrated approach to extracting ontological structures from folksonomies. In *6th Annual European Semantic Web Conference (ESWC2009)*, pages 654–668, June 2009.

[77] Vanessa Lopez, Andriy Nikolov, Marta Sabou, Victoria Uren, and Enrico Motta. Scaling up question-answering to linked data, knowledge engineering and knowledge management by the masses. In *Proceedings of the Conference for Knowledge Engineering and Knowledge Management by the Masses (EKAW)*, 2010.

[78] Mathias Lux, Michael Granitzer, and Roman Kern. Aspects of broad folksonomies. In *Database and Expert Systems Applications, 2007. DEXA '07. 18th International Conference on*, pages 283–287, 2007.

[79] Mohamed Zied Maala, Alexandre Delteil, and Ahmed Azough. A conversion process from flickr tags to rdf descriptions. In *Proceedings of the 10th International Conference on Business Information Systems*, Poznan, Poland, 2007.

[80] Alexander Maedche and Steffen Staab. Measuring similarity between ontologies. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*, pages 251–263, 2002.

[81] Ching man Yeung Au, Nicholas Gibbins, and Nigel Shadbolt. A k-nearest-neighbour method for classifying web search results with data in folksonomies. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, pages 70–76, Washington, DC, USA, 2008. IEEE Computer Society.

[82] Andrea Marchetti, Maurizio Tesconi, and Francesco Ronzano. Semkey: A semantic collaborative tagging system. In *WWW07 Workshop, Tagging and Metadata for Social Information Organization*, 2007.

[83] Cameron Marlow, Mor Naaman, Danah Boyd, and Marc Davis. Position paper, tagging, taxonomy, flickr, article, toread. In *Collaborative Web Tagging Workshop (WWW '06)*, 2006.

[84] Adam Mathes. Folksonomies - cooperative classification and communication through shared metadata.

[85] Pasquale De Meo, Giovanni Quattrone, and Domenico Ursino. A query expansion and user profile enrichment approach to improve the performance of recommender systems operating on a folksonomy. *User Modeling and User-Adapted Interaction*, 20(1):41–86, 2010.

[86] Peter Mika. Ontologies are us: A unified model of social networks and semantics. In *Proceedings of the 4th International Semantic Web Conference*, pages 522–536. Springer SE - LNCS, 2005.

[87] David Millen, Meng Yang, Steven Whittaker, and Jonathan Feinberg. Social bookmarking and exploratory search. In *ECSCW 2007*, pages 40, 21. 2007.

[88] Alan Mislove, Krishna P. Gummadi, and Peter Druschel. Exploiting social networks for internet search. In *Proceedings of the 5th Workshop on Hot Topics in Networks (HotNets'06)*, 2006.

[89] Roberto Navigli and Paola Velardi. An analysis of ontology-based query expansion strategies. In *Workshop on Adaptive Text Extraction and Mining, (Cavtat Dubrovnik, Croatia, Sept 23)*, 2003.

[90] Richard Newman, Danny Ayers, and Seth Russell. Tag ontology, 2005.

[91] Enkhbold Nyamsuren and Ho-Jin Choi. Building domain independent ontology for web 2.0. In *Computer and Information Technology Workshops, 2008. CIT Workshops 2008. IEEE 8th International Conference on*, pages 655–660, 2008.

[92] Tim O'Reilly. What is web2.0?, September 2005.

[93] Eyal Oren, A. Haller, M. Hauswirth, B. Heitmann, S. Decker, and C. Mesnage. A flexible integration framework for semantic Web 2.0 applications. *Software, IEEE*, 24:64–71, 2007.

[94] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab, 1999.

[95] Jeff Z. Pan, Steve Taylor, and Edward Thomas. Expanding folksonomy search with ontologies. In Christian Bizer and Anupam Joshi, editors, *Proceedings of the 7th International Semantic Web Conference (Posters & Demos)*, volume 401 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[96] Jeff Z. Pan, Stuart Taylor, and Edward Thomas. Reducing ambiguity in tagging systems with folksonomy search expansion. In *Proceedings of the 6th European Semantic Web Conference on The Semantic Web: Research and Applications*, pages 669–683, Heraklion, Crete, Greece, 2009. Springer-Verlag.

[97] Alexandre Passant. Using ontologies to strengthen folksonomies and enrich information retrieval in weblogs: Theoretical background and corporate use-case. In *ICWSM 2007*.

[98] Alexandre Passant and Phillippe Laublet. Meaning of a tag: A collaborative approach to bridge the gap between tagging and linked data. In *Proceedings of the WWW 2008 Workshop Linked Data on the Web (LDOW2008), Beijing, China, Apr*, 2008.

[99] Elaine Peterson. Beneath the metadata: Some philosophical problems with folksonomy. D-Lib Magazine, 2006.

[100] Anon Plangprasopchok and Kristina Lerman. Modeling social annotation: a bayesian approach. *CoRR*, abs/0811.1319, 2008.

[101] Anon Plangprasopchok and Kristina Lerman. Constructing folksonomies from user-specified relations on flickr. In *Proceedings of the 18th International Conference on World wide web*, pages 781–790, New York, NY, USA, 2009.

[102] Anon Plangprasopchok, Kristina Lerman, and Lise Getoor. Constructing folksonomies by integrating structured metadata. In *WWW '10: Proceedings of the 19th international conference on World wide web*, pages 1165–1166, New York, NY, USA, 2010. ACM.

[103] Emanuele Quintarelli, Andrea Resmini, and Luca Rosati. Facetag: Integrating bottom-up and top-down classification in a social tagging system. In *iA SUMMIT*, Las Vegas, USA, 2007.

[104] Raghu Ramakrishnan and Andrew Tomkins. Toward a PeopleWeb. *Computer*, 40(8):63–72, 2007.

[105] Tye Rattenbury, Nathaniel Good, and Mor Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 103–110, New York, NY, USA, 2007. ACM.

[106] Francesco Ronzano, Andrea Marchetti, and Maurizio Tesconi. Tagpedia: a semantic reference to describe and search for web resources. In Peter Dolog, Markus Krtzsch, Sebastian Schaffert, and Denny Vrandecic, editors, *SWKM*, volume 356 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2008.

[107] Terrell Russell. Cloudalicious: folksonomy over time. In *Digital Libraries, 2006. JCDL '06. Proceedings of the 6th ACM/IEEE-CS Joint Conference on*, page 364, 2006.

[108] Marta Sabou, Mathieu d'Aquin, and Enrico Motta. Using the semantic web as background knowledge for ontology mapping. In *International Workshop on Ontology Matching (OM-2006), collocated with ISWC'06*, Athens, Georgia, USA, 2006.

[109] Patrick Schmitz. Inducing ontology from flickr tags. In *Collaborative Web Tagging Workshop (WWW '06)*, 2006. TY - CONF.

[110] Nigel Shadbolt, Tim Berners-Lee, and Wendy Hall. The semantic web revisited. *IEEE Intelligent Systems*, 21:96–101, 2006.

[111] Hana Shepard, Harry Halpin, and Valentin Robu. The dynamics and semantics of collaborative tagging. In *Proceedings of the 1st Semantic Authoring and Annotation Workshop*, 2006.

[112] Clay Shirky. Ontology is overrated: Categories, links, and tags, 2005.

[113] Katharina Siorpaes. myontology: The marriage of ontology engineering and collective intelligence. In *Proceedings of the ESWC workshop: Bridging the Gap between Semantic Web and Web 2.0*, pages 127–138, 2007.

[114] Lucia Specia and Enrico Motta. Integrating folksonomies with the semantic web. In *Proceedings of the 4th ESWC*, pages 624–639, 2007.

[115] Martin Szomszor, Harith Alani, Ivan Cantador, Kieron O'Hara, and Nigel Shadbolt. Semantic modelling of user interests based on cross-folksonomy analysis.

In *Proceedings of the 7th International Conference on The Semantic Web*, pages 632–648, 2008.

[116] Maurizio Tesconi, Francesco Ronzano, Andrea Marchetti, and Salvatore Minutoli. Semantify del.icio.us: automatically turn tags into senses. In *1st workshop on Social Data on the Web (SDow2008)*, October 2008.

[117] Raquel Trillo, Jorge Gracia, Mauricio Espinoza, and Eduardo Mena. Discovering the semantics of user keywords. *Journal of Universal Computer Science*, 13(12):1908–1935, 2007.

[118] Karen Tso-Sutter, Leandro Marinho, and Lars Schmidt-Thieme. Tag-aware recommender systems by fusion of collaborative filtering algorithms. In *Proceedings of the 23rd ACM Symposium on Applied Computing (SAC 2008)*, pages 1995–1999, New York, NY, USA, March 2008. ACM.

[119] Reinier H. van Leuken, Lluis Garcia, Ximena Olivares, and Roelof van Zwol. Visual diversification of image search results. In *WWW '09: Proceedings of the 18th international conference on World wide web*, pages 341–350, New York, NY, USA, 2009. ACM.

[120] Roelof van Zwol. Flickr: Who is looking? In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 184–190. IEEE Computer Society, 2007.

[121] Roelof van Zwol, Vanessa Murdock, Lluis Garcia Pueyo, and Georgina Ramirez. Diversifying image search with user generated content. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 67–74, New York, NY, USA, 2008. ACM.

[122] Thomas Vanderwal. Folksonomy coinage and definition, February 2007.

[123] David Weinberger. Tagging and why it matters. Berkman Center Research Publication, 2005.

[124] Xian Wu, Lei Zhang, and Yong Yu. Exploring social annotations for the semantic web. In *Proceedings of the 15th international conference on World Wide Web*, pages 417–426. ACM, 2006.

[125] Zhibiao Wu and Martha Palmer. Verb semantics and lexical selection. In *Proceedings of the 32nd Annual Meeting of the Association for Computational Linguistics*, pages 133 –138, New Mexico, USA, 1994.

[126] Sihem A. Yahia, Alban Galland, Julia Stoyanovich, and Cong Yu. From del.icio.us to x.qui.site: recommendations in social tagging sites. In Jason Tsong Li Wang and Jason Tsong Li Wang, editors, *SIGMOD Conference*, pages 1323–1326, New York, NY, USA, 2008. ACM.

[127] Yusuke Yanbe, Adam Jatowt, Satoshi Nakamura, and Katsumi Tanaka. Towards improving web search by utilizing social bookmarks. In *Proceedings of the 7th international conference on Web engineering*, pages 343–357, 2007.

[128] Ching-man A. Yeung, Nicholas Gibbins, and Nigel Shadbolt. Understanding the semantics of ambiguous tags in folksonomies. In *International Semantic Web Conference*, Busan, South Korea, 2007.

[129] Ching-man A. Yeung, Nicholas Gibbins, and Nigel Shadbolt. Contextualising tags in collaborative tagging systems. In *HT '09: Proceedings of the 20th ACM conference on Hypertext and hypermedia*, pages 251–260, New York, NY, USA, 2009. ACM.

[130] Ching-man A. Yeung, Nicholas Gibbins, and Nigel Shadbolt. User-induced links in collaborative tagging systems. In *CIKM '09: Proceeding of the 18th ACM conference on Information and knowledge management*, pages 787–796, New York, NY, USA, 2009. ACM.

[131] Donghee Yoo and Yongmoo Suh. User-categorized tags to build a structured folksonomy. *Communication Software and Networks, International Conference on*, 0:160–164, 2010.

[132] Fouad Zablith, Mathieu d'Aquin, Marta Sabou, and Enrico Motta. Using ontological contexts to assess the relevance of statements in ontology evolution. In *EKAW*, pages 226–240, 2010.

[133] Valentina Zanardi and Licia Capra. Social ranking: uncovering relevant content using tag-based recommender systems. In *Proceedings of the 2008 ACM Conference on Recommender Systems*, pages 51–58, New York, USA, 2008.

[134] Shiwan Zhao, Nan Du, Andreas Nauerz, Xiatian Zhang, Quan Yuan, and Rongyao Fu. Improved recommendation based on collaborative tagging behaviors. In *Proceedings of the 13th International Conference on Intelligent User Interfaces*, pages 413–416, New York, USA, 2008.

[135] Mianwei Zhou, Shenghua Bao, Xian Wu, and Yong Yu. An unsupervised model for exploring hierarchical semantics from social annotations. In *Proceedings of the 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference*, pages 680–693, Seoul, Korea, 2007.

[136] Alla Zollers. Emerging motivations for tagging: Expression, performance, and activism. In *Proceedings of the 16th International Conference on World Wide Web*, Banff, Canada, 2007.